دانشکده.....

گزارش .....

عنوان :.....

ارائه‌دهندگان:

.....

استاد درس:

.....

زمستان ۱۴۰۰


بسم الله الرحمن الرحیم

چکیده

تشخیص حالت چهره در ۱۰ سال گذشته یک حوزه تحقیقاتی فعال بوده است، حوزه‌های کاربردی رو به رشدی ازجمله انیمیشن آواتار، بازاریابی عصبی و روبات‌های اجتماعی در این حوزه فعالیت داشتند. تشخیص حالات چهره روش آسانی برای روش‌های یادگیری ماشینی نیست، زیرا افراد می‌توانند به‌طور قابل‌توجهی در نحوه نشان دادن حالات خود متفاوت باشند. حتی تصاویر یک فرد در حالت چهره یکسان می‌تواند درروشنایی، پس‌زمینه و ژست متفاوت باشد، و این تغییرات با در نظر گرفتن موضوعات مختلف (به دلیل تفاوت در شکل و قومیت) نیز وجود دارد. اگرچه تشخیص حالت چهره بسیار موردمطالعه قرارگرفته است، اما تعداد کمی از آن‌ها ارزیابی درستی را انجام می‌دهند و از مشکلات در حین آموزش و آزمایش الگوریتم‌های پیشنهادی اجتناب می‌کنند. ازاین‌رو، تشخیص حالت چهره هنوز یک مشکل چالش‌برانگیز در بینایی کامپیوتر است. در این مطالعه، ما یک‌راه حل ساده برای تشخیص حالت چهره ارائه می‌کنیم که از ترکیب شبکه عصبی کانولوشنال و مراحل پیش‌پردازش تصویر استفاده می‌کند. شبکه‌های عصبی کانولوشنال با داده‌های بزرگ به‌دقت بهتری دست می‌یابند. بااین‌حال، هیچ مجموعه داده در دسترس با داده‌های کافی برای تشخیص حالت چهره با یادگیری عمیق وجود ندارد. بنابراین، برای مقابله با این مشکل، برخی از تکنیک‌های پیش‌پردازش را به کار می‌بریم تا فقط ویژگی‌های خاص را از تصویر چهره استخراج‌کنیم و ترتیب ارائه نمونه‌ها را در طول آموزش بررسی کنیم. آزمایش‌های مورداستفاده برای ارزیابی تکنیک با استفاده از سه پایگاه داده عمومی (CKþ، JAFFE و BU-3DFE) انجام شد. روش پیشنهادی، در مقایسه با سایر روش‌های تشخیص حالات چهره نتایج قابل قبولی را به دست می‌آورد. ۹۶/۷۶ درصد دقت در پایگاه داده CKþ به‌دست‌آمده است. این آموزش سریع است و امکان تشخیص حالت چهره در زمان واقعی را با رایانه‌های استاندارد فراهم می‌کند.

فهرست مطالب

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست علایم و نشانه‌ها

عنوان	علامت اختصاری
تعداد نمونه‌ها	N
تعداد حالت‌ها	T
پیکسل جدید	X'
پیکسل قبلی	X

۱– معرفی

حالت چهره یکی از مهم‌ترین ویژگی‌های تشخیص احساسات انسان است [۱]. داروین در کتاب «بیان عواطف در انسان و حیوانات»[۲] به‌عنوان یک زمینه تحقیقاتی به تشخیص چهره اشاره می‌کند. به گفته لی و جین [۳]، می‌توان آن را به‌عنوان تغییرات چهره در پاسخ به وضعیت عاطفی درونی، نیات یا ارتباطات اجتماعی فرد تعریف کرد. امروزه، تشخیص خودکار حالت چهره دارای کاربردهای متنوعی است، مانند انیمیشن مبتنی بر داده[۱]، بازاریابی عصبی[۲]، بازی‌های تعاملی[۳]، روباتیک اجتماعی [۴]و بسیاری دیگر از سیستم‌های تعامل انسان و رایانه.

تشخیص چهره وظیفه‌ای است که انسان به‌صورت روزانه و بدون زحمت انجام می‌دهد [۳]، اما باوجوداینکه روش‌های اخیر در برخی شرایط با دقت بیش از ۹۵ درصد ارائه‌شده است، هنوز توسط رایانه‌ها به‌راحتی انجام نمی‌شود (چهره، محیط‌های کنترل‌شده و تصاویر با وضوح‌بالا). بسیاری از آثار موجود، یک روش ارزیابی منسجم را انجام نمی‌دهند (مثلاً بدون همپوشانی در آموزش و آزمایش) و بنابراین دقت بالایی را ارائه نمی‌کنند و بیشتر مشکلات تشخیص چهره را نشان نمی‌دهند. از سوی دیگر، دقت پایینی در پایگاه‌های اطلاعاتی با محیط‌های کنترل‌شده و در ارزیابی‌های پایگاه داده گزارش‌شده است. چندین کار تحقیقاتی سعی کرده‌اند رایانه‌ها را به‌دقت انسان‌ها برسانند، و نمونه‌هایی از این کارها نام‌برده شده‌اند. این مشکل هنوز برای کامپیوترها چالش‌برانگیز است زیرا تفکیک فضای ویژگی عبارات بسیار سخت است، یعنی ویژگی‌های صورت در دو حالت مختلف ممکن است در فضای ویژگی بسیار نزدیک باشد، درحالی‌که ویژگی‌های چهره باحالت یکسان ممکن است بسیار نزدیک باشد یا ممکن است از یکدیگر بسیار دور باشند. علاوه بر این، برخی از حالات مانند غم و ترس، در برخی موارد بسیار شبیه هستند.

شکل ۱ سه موضوع را با چهره شاد نشان می‌دهد. همان‌طور که در شکل مشاهده می‌شود، تصاویر نه‌تنها در نحوه بیان سوژه‌ها، بلکه در نور، روشنایی، ژست و پس‌زمینه نیز با یکدیگر تفاوت زیادی دارند. این شکل همچنین چالش دیگری در رابطه با تشخیص حالت چهره را نشان می‌دهد که سناریوهای آزمایش-آموزش کنترل نشده است (تصاویر آموزشی می‌تواند ازنظر شرایط محیطی با تصاویر آزمایشی بسیار متفاوت باشد). یک رویکرد برای ارزیابی تشخیص حالت چهره، آموزش روش با یک پایگاه داده و آزمایش آن با دیگری (احتمالاً از گروه‌های قومی مختلف) است.

---

[1] Data-driven animation
[2] Neuromarketing
[3] Interactive games
[4] Sociable robotics

شکل ۱: سه حالت مختلف شاد. تصاویر از پایگاه‌های داده CKþ، پایگاه داده JAFF و پایگاه داده BU-3DFE.

سیستم‌های تشخیص حالت چهره را می‌توان به دودسته اصلی تقسیم کرد: آن‌هایی که با تصاویر ثابت [۷-۱۳] کار می‌کنند و آن‌هایی که با دنباله‌های تصویر پویا [۱۴-۱۷][1] کار می‌کنند. روش‌های مبتنی بر استاتیک[2] از اطلاعات زمانی استفاده نمی‌کنند، یعنی بردار ویژگی فقط شامل اطلاعات مربوط به تصویر ورودی فعلی است. از سوی دیگر، روش‌های مبتنی بر توالی، از اطلاعات زمانی تصاویر برای تشخیص استفاده می‌کنند. سیستم‌های خودکار برای تشخیص حالات چهره ورودی مورد انتظار (تصویر ثابت یا دنباله تصویر) را دریافت می‌کنند و معمولاً یکی از شش حالت اصلی (برای مثال خشم، غم، تعجب، خوشحالی، انزجار و ترس) را به‌عنوان خروجی ارائه می‌دهند. برخی از دستگاه‌ها نیز حالت خنثی را تشخیص می‌دهند. این کار بر روی روش‌های مبتنی بر تصاویر ثابت تمرکز می‌کند و مجموعه هفت حالت (شش حالت پایه به‌علاوه خنثی) را برای سناریوهای کنترل‌شده و کنترل‌شده در نظر می‌گیرد.

همان‌طور که توسط لی و جین [۳] توضیح داده شد، تجزیه‌وتحلیل خودکار حالت چهره شامل سه مرحله است: انتخاب چهره، استخراج و نمایش داده‌های چهره، و تشخیص حالت چهره. انتخاب چهره را می‌توان به دو مرحله عمده تقسیم کرد: تشخیص چهره [۱۸-۲۱] و تخمین وضعیت سر [۲۲-۲۴]. پس از گرفتن چهره، تغییرات صورت ناشی از حالات چهره باید استخراج شود. این تغییرات معمولاً با استفاده از روش‌های مبتنی بر ویژگی هندسی [۲۵-۲۱،۲۷] یا روش‌های مبتنی بر ظاهر[3] [۸،۲۵-۱۳،۲۸،۱۱] استخراج می‌شوند. ویژگی‌های استخراج‌شده اغلب در بردارها نشان داده می‌شوند که به‌عنوان بردارهای ویژگی نامیده می‌شوند. روش‌های مبتنی بر ویژگی‌های هندسی با شکل و مکان اجزای صورت مانند دهان، چشم‌ها، بینی و ابروها کار می‌کنند. بردار ویژگی که هندسه چهره را نشان می‌دهد از اجزای صورت یا نقاط ویژگی صورت تشکیل‌شده است. روش‌های مبتنی بر ظاهر با بردارهای ویژگی استخراج‌شده از کل

[1] Image sequences
[2] Static-based methods
[3] Appearance-based methods

صورت، یا از مناطق خاص کار می‌کنند. این بردارهای ویژگی با استفاده از فیلترهای تصویر اعمال‌شده بر روی کل تصویر به دست می‌آیند [۳].

هنگامی‌که بردارهای ویژگی مربوط به حالت چهره در دسترس هستند، می‌توان تشخیص حالت را انجام داد. به گفته لئو و همکاران [۷]، سیستم‌های تشخیص حالت اساساً از یک روش آموزشی سه مرحله‌ای استفاده می‌کنند، یادگیری ویژگی، انتخاب ویژگی و ساخت طبقه‌بندی کننده. مرحله یادگیری ویژگی مسئول استخراج تمام ویژگی‌های مربوط به حالت چهره است. انتخاب ویژگی، بهترین ویژگی‌ها را برای نمایش مجدد حالت چهره انتخاب می‌کند. آن‌ها باید تنوع درون کلاسی را به حداقل برسانند درحالی‌که تنوع بین کلاسی را به حداکثر برسانند [۸]. به حداقل رساندن تنوع درون کلاسی یک مشکل است زیرا تصاویر افراد مختلف باحالت مشابه در فضای پیکسل از یکدیگر دور هستند. حداکثر رساندن تنوع بین کلاسی نیز دشوار است زیرا تصاویر یک فرد در حالات مختلف ممکن است در فضای پیکسل بسیار نزدیک به یکدیگر باشند [۲۹]. در پایان کل فرآیند، یک طبقه‌بندی (یا مجموعه‌ای از طبقه‌بندی) برای استنباط حالت چهره، با توجه به ویژگی‌های انتخاب‌شده، استفاده می‌شود.

یکی از تکنیک‌هایی که به‌طور موفقیت‌آمیزی برای مشکل تشخیص حالت چهره به کار گرفته‌شده است، شبکه عصبی چندلایه عمیق [۳۰-۳۲،۱۴،۱۰،۱۱،۱۳] بود. این تکنیک شامل سه مرحله تشخیص حالات چهره (یادگیری و انتخاب ویژگی‌ها و طبقه‌بندی) در یک مرحله است. در دهه گذشته، تحقیقات شبکه‌های عصبی باانگیزه یافتن راهی برای آموزش شبکه‌های عصبی چندلایه عمیق (یعنی شبکه‌هایی با بیش از یک یا دولایه پنهان) به‌منظور افزایش دقت آن‌ها بود [۳۳،۳۴]. طبق گفته Bengio [۳۵]، تا سال ۲۰۰۶، بسیاری از تلاش‌های جدید موفقیت چندانی از خود نشان نداده‌اند. اگرچه تا حدودی قدیمی است، اما شبکه‌های عصبی کانولوشنال [1](CNN) در سال ۱۹۹۸ توسط Lecun و همکاران [۳۶] نشان داده است که در یادگیری ویژگی‌ها، هنگام استفاده از معماری‌های عمیق‌تر (یعنی با لایه‌های زیاد) و تکنیک‌های آموزشی جدید بسیار مؤثر است. به‌طورکلی، این نوع شبکه دارای انواع لایه‌های متناوب شامل لایه‌های کانولوشن، لایه‌های نمونه فرعی و لایه‌های کاملاً متصل است. لایه‌های کانولوشن با اندازه هسته مشخص می‌شوند. لایه‌های نمونه‌گیری فرعی برای افزایش تغییرناپذیری موقعیت هسته‌ها با کاهش اندازه نقشه استفاده می‌شوند [۳۷]. لایه‌های کاملاً متصل در شبکه‌های عصبی CNN مشابه لایه‌های شبکه‌های عصبی عمومی هستند [۳۷]، نورون‌های آن کاملاً با لایه قبلی (به‌طورکلی لایه کانوولوشن، لایه زیر نمونه‌برداری یا حتی یک‌لایه کاملاً متصل) مرتبط هستند. یادگیری نظارت‌شده را می‌توان با استفاده از روش نزولی گرادیان انجام داد، مانند روشی که توسط Lecun و همکاران ارائه شد [۳۶]. یکی از

[1] Convolu- tional Neural Networks

مزیت‌های اصلی CNN این است که ورودی مدل‌ها یک تصویر خام است نه مجموعه‌ای از ویژگی‌های کدگذاری شده دستی[1].

علاوه بر روش‌هایی که از معماری عمیق استفاده می‌کنند، روش‌های بسیار دیگری نیز وجود دارد، اما برخی از جنبه‌های ارزیابی این روش‌ها هنوز هم موردتوجه است. برای مثال، روش‌های اعتبارسنجی را می‌توان در [۳۸-۴۲،۳۰،۱۰] بهبود بخشید، دقت تا حدودی در [۴۴،۴۳،۱،۳۸]، و زمان تشخیص در [۱۶،۴۳،۷] می‌تواند به‌گونه‌ای بهبود یابد که ارزیابی‌های زمان واقعی انجام شود.

در تلاش برای مقابله با برخی از این محدودیت‌ها و درعین‌حال حفظ یک راه‌حل ساده، در این مقاله، یک رویکرد یادگیری عمیق را ارائه می‌کنیم که روش‌های استاندارد را ترکیب می‌کند، مانند نرمال‌سازی تصویر[2]، تولید نمونه‌های آموزشی مصنوعی (به‌عنوان‌مثال، تصاویر واقعی با چرخش مصنوعی[3]، انتقال[4] و مقیاس بندی[5]) و شبکه عصبی کانولوشن، به یک‌راه حل ساده که قادر به دستیابی به نرخ دقت بسیار بالای ۹۶/۹۷ درصد در پایگاه داده CKþ برای شش حالت است. زمان آموزش در مقایسه با روش‌های دیگر به‌طور قابل‌توجهی کمتر است و کل سیستم تشخیص حالت چهره می‌تواند در زمان واقعی در رایانه‌های استاندارد کار کند. ما عملکرد سیستم خود را با استفاده از پایگاه داده گسترده Cohn-Kanade (CKþ)، پایگاه داده بیان چهره زنان ژاپنی (JAFFE) و پایگاه داده بیان چهره سه‌بعدی دانشگاه بینگامتون (BU-3DFE) بررسی کرده‌ایم.، دستیابی به‌دقت بهتر در پایگاه داده CKþ که شامل نمونه‌های بیشتری (برای تکنیک‌های یادگیری عمیق مهم است) نسبت به پایگاه‌های داده JAFFE و BU-3DFE است. علاوه بر این، ما اعتبار سنجی گسترده‌ای را با آزمایش‌های متقابل پایگاه داده (یعنی آموزش با استفاده از یک پایگاه داده و ارزیابی دقت آن با استفاده از پایگاه داده دیگر) انجام داده‌ایم. به‌طور خلاصه، کارهای اصلی این مقاله عبارت‌اند از:

۱- یک روش کارآمد برای تشخیص حالات چهره که در زمان واقعی عمل می‌کند.
۲- مطالعه اثرات عملیات پیش‌پردازش تصویر در مشکل تشخیص حالت چهره.
۳- مجموعه‌ای از عملیات پیش‌پردازش برای عادی‌سازی چهره (شدت روشنایی) به‌منظور کاهش نیاز به محیط‌های کنترل‌شده و مقابله با کمبود داده.
۴- مطالعه‌ای برای رسیدگی به‌دقت ناشی از ترتیب ارائه نمونه‌ها در طول آموزش؛ و
۵- مطالعه عملکرد سیستم پیشنهادی با محیط‌های مختلف (ارزیابی متقابل پایگاه داده).

[1] Hand-coded features
[2] Image normalizations
[3] Artificial rotations
[4] Translation
[5] Scaling

۶- این اثر ارائه‌شده در بیست و هشتمین کنفرانس SIBGRAPI (کنفرانس گرافیک، الگوها و تصاویر) [۴۵] را به شرح زیر بیان می‌کند:

۷- این مقاله مرور عمیق‌تری را ارائه می‌کند که برای نشان دادن مقایسه‌های نتایج استفاده‌شده است.

۸- نتایج را با استفاده از یک پیاده‌سازی جدید ارائه می‌دهد

۹- چارچوب متفاوت، که درنتیجه کل زمان آموزش را تقریباً چهار برابر کاهش داد.

۱۰- نتایجی را نشان می‌دهد که زمان شناسایی کاهش‌یافته را نشان می‌دهد.

۱۱- نتایجی را ارائه می‌دهد که دقت بهتری را به دلیل آموزش طولانی‌تر و تغییرات کوچک درروش نشان می‌دهد.

۱۲- روش تجربی بهبودیافته است، به‌عنوان‌مثال، استفاده از آموزش، اعتبار سنجی و مجموعه‌های آزمایشی، به‌جای آموزش و مجموعه‌های آزمایشی.

۱۳- یک ارزیابی کامل‌تر از آزمون‌های بین پایگاه داده ارائه می‌دهد.

تغییراتی که امکان دقت بهتر را فراهم کرد به شرح زیر بود:

i. نمونه‌های مصنوعی فرآیند تولید متفاوت دارند، که امکان تنوع بیشتر در بین آن‌ها را فراهم می‌کند (اکنون نمونه‌های مصنوعی می‌توانند به‌جای چرخاندن تصویر اصلی، چرخانده یا مقیاس بندی یا منتقل شوند).

ii. ما تعداد نمونه‌های مصنوعی را از ۳۰ به ۷۰ افزایش دادیم.

iii. تابع تلفات رگرسیون لجستیک [1]با یک تابع Soft maxWithLoss جایگزین شد.

بقیه مقاله به شرح زیر سازمان‌دهی شده است، بخش بعدی جدیدترین کار مرتبط را ارائه می‌دهد، و بخش ۳ رویکرد پیشنهادی را تشریح می‌کند. در بخش ۴، آزمایش‌هایی که برای ارزیابی سیستم خود انجام داده‌ایم، ارائه‌شده و با چندین روش اخیر تشخیص چهره مقایسه شده‌اند. درنهایت در بخش ۵ نتیجه می‌گیریم.

۲- کارهای مرتبط

چندین رویکرد تشخیص حالت چهره در دهه‌های گذشته با پیشرفت روزافزون توسعه‌یافته است. بخش مهمی از این پیشرفت اخیر به خاطر روش‌های یادگیری عمیق [۱۲،۱۰،۷] و به‌طور خاص‌تر با شبکه‌های عصبی کانولوشن [۱۱،۱۴] که یکی از رویکردهای یادگیری عمیق است، به دست آمد. این

[1] Logistic regression loss function

رویکردها به دلیل حجم بیشتر داده‌های موجود برای آموزش، روش‌های یادگیری و پیشرفت در فناوری GPU امکان‌پذیر شد. که برای شبکه‌های آموزشی با معماری‌های عمیق بسیار مهم است، همچنین دومی برای محاسبات عددی کم‌هزینه و با کارایی بالا برای روش آموزشی بسیار مهم است. نظرسنجی از تحقیقات تشخیص حالت چهره را می‌توان در [۴۸،۳] یافت.

برخی از رویکردهای اخیر برای تشخیص حالات چهره بر روی محیط‌های کنترل نشده متمرکزشده‌اند (به‌عنوان‌مثال چهره از روبرو، تصاویری که همپوشانی دارند، حالات غیرارادی)، که هنوز یک مشکل چالش‌برانگیز است [۵۰،۴۹،۱۲]. این کار بر روی محیط‌های کنترل‌شده و ارزیابی در میان گروه‌های قومی مختلف تمرکز خواهد کرد. این بخش روش‌های اخیری را موردبحث قرار می‌دهد که با استفاده از روش‌شناسی تجربی قابل‌مقایسه یا روش‌هایی که مبتنی بر شبکه‌های عصبی عمیق هستند، به‌دقت بالایی در تشخیص حالت چهره دست می‌یابند.

لئو و همکاران [۷] رویکرد جدیدی به نام شبکه باور عمیق تقویت‌شده (BDBN)[1] پیشنهاد کرد. BDBN توسط مجموعه‌ای از طبقه‌بندی کننده‌ها تشکیل‌شده است که توسط نویسندگان به‌عنوان طبقه‌بندی کننده ضعیف نام‌گذاری شده است. هر طبقه‌بندی کننده ضعیف مسئول طبقه‌بندی یک حالت است. رویکرد آن‌ها سه مرحله یادگیری (یادگیری ویژگی، انتخاب ویژگی و ساخت طبقه‌بندی کننده) را به‌طور مکرر در یک چارچوب منحصربه‌فرد انجام می‌دهد. آزمایش‌های آن‌ها با استفاده از دو پایگاه داده عمومی تصاویر ثابت، Cohn-Kanade و JAFFE [۵۱] انجام شد و به ترتیب به‌دقت ۹۶/۷ و ۹۱/۸ درصد رسید. همچنین آزمایش‌هایی را روی سناریوهای کمتر کنترل‌شده بین پایگاه داده (آموزش با CKþ و آزمایش در JAFFE) انجام دادند و به‌دقت ۸۶/۰ درصد دست یافتند. همه تصاویر ابتدا بر اساس مختصات چشمی[2] داده‌شده. آموزش و آزمون روش طبقه‌بندی یک در مقابل همه [3]را اتخاذ کردند، یعنی از یک طبقه‌بندی کننده باینری برای هر حالت استفاده شد. مدت‌زمان لازم برای آموزش شبکه حدود ۸ روز بود. تشخیص به‌عنوان تابعی از طبقه‌بندی کننده‌های ضعیف محاسبه شد. آن‌ها از شش یا هفت طبقه‌بندی کننده، بسته به میزان حالاتی که باید شناسایی شوند، استفاده می‌کنند. هر طبقه‌بندی کننده ۳۰ میلی‌ثانیه طول می‌کشد تا هر حالت را تشخیص دهد، زمان کل تشخیص حدود ۰/۲۱ ثانیه شد. زمان تشخیص توسط نویسندگان با استفاده از یک رایانه شخصی شش هسته‌ای ۲/۴ گزارش‌شده است.

[1] Boosted deep belief network
[2] Eye coordinates
[3] One-versus-al

سونگ و همکاران [۱۰]، یک سیستم تشخیص حالت چهره را توسعه دادند که از یک شبکه عصبی کانولوشنال عمیق استفاده می‌کند و بر روی تلفن هوشمند اجرا می‌شود. شبکه پیشنهادی از پنج لایه و ۶۵۰۰۰ نورون تشکیل‌شده است. نویسندگان بعدی از تکنیک‌های افزایش داده‌ها برای افزایش میزان داده‌های آموزشی و از حذف [۵۲] در طول آموزش شبکه استفاده کردند. آزمایش‌ها با استفاده از مجموعه داده CKþ و سه مجموعه داده دیگر انجام شد. تصاویر مجموعه داده CKþ ابتدا برای تمرکز بر روی مناطقی که حاوی تغییرات صورت ناشی از یک حالت هستند، برش داده شدند. آزمایش‌های انجام‌شده توسط نویسندگان از اعتبارسنجی متقابل ۱۰ برابری [1]پیروی می‌کند، اما آن‌ها اشاره نمی‌کنند که تصاویری از یک حالت وجود داشته است یا خیر. بنابراین، ما فرض کردیم که بین مجموعه‌های آموزشی و آزمون همپوشانی وجود دارد. دقت ۹۹/۲ درصد در پایگاه داده CKþ به دست آمد درحالی‌که فقط پنج حالت (خشم، خوشحالی، غم، تعجب و خنثی) را تشخیص داد.

برکرت و همکاران [۱۱] روشی مبتنی بر شبکه‌های عصبی کانولوشنال پیشنهاد کردند. نویسندگان ادعا می‌کنند که روش آن‌ها مستقل از هرگونه استخراج ویژگی دست‌ساز است (یعنی از تصویر خام به‌عنوان ورودی استفاده می‌کند). معماری شبکه آن‌ها از چهار جز تشکیل‌شده است

بخش اول وظیفه پیش‌پردازش خودکار داده‌ها را بر عهده دارد، درحالی‌که بخش‌های دیگر فرآیند استخراج ویژگی را انجام می‌دهند. ویژگی‌های استخراج‌شده توسط یک‌لایه کاملاً متصل در انتهای شبکه به یک حالت طبقه‌بندی می‌شوند. معماری پیشنهادی شامل ۱۵ لایه (۷ کانولوشن، ۵ ادغام، ۱ لایه نرمال‌سازی) است. روش خود را با پایگاه داده CKþ و MMI da tabase ارزیابی کردند و به ترتیب به‌دقت ۹۹/۶ درصد و ۹۸/۶۸ درصد دست یافتند. علی رقم دقت بالا به همپوشانی نداشتن داده آموزش و آزمایش دقت نکردند. همان‌طور که در بخش ۴ موردبحث قرار خواهد گرفت، این یک محدودیت مهم است که باید به‌منظور انجام یک ارزیابی خوب از روش‌های تشخیص حالت چهره اعمال شود [۵۳،۴۴].

لئو و همکاران [۱۲] از شبکه‌های عمیق الهام گرفتند و واحد عمل (AUDNوAU) را پیشنهاد کردند و یک نظریه روان‌شناختی را بررسی کردند. نویسندگان ادعا می‌کنند که این روش قادر به یادگیری، تغییرات ظاهری محلی آموزنده و یک‌راه بهینه برای ترکیب تغییرات محلی و یک نمایش سطح بالا برای تشخیص عبارت نهایی است. آزمایش‌ها در مجموعه داده‌های CKþ، MMI [۵۴] و SFEW [۵۵] انجام شد. MMI شامل تصاویر گرفته‌شده از فیلم‌های مختلف تحت سناریوهای کنترل نشده است که نشان‌دهنده محیط دنیای واقعی است. آزمایش‌ها با استفاده از رویکرد اعتبارسنجی متقابل بدون همپوشانی موضوعی بین گروه‌های آموزشی و آزمون، و ارزیابی شش حالت اصلی شکل گرفتند. این روش به‌دقت ۹۳/۷۰

[1] 10-fold cross validation

درصد در پایگاه داده CKþ، ۷۵/۸۵ درصد در پایگاه داده MMI و ۳۰/۱۴ درصد در پایگاه داده SFEW دست می‌یابد.

علای و همکاران [۱۳] مجموعه‌ای از شبکه عصبی تقویت‌شده برای تشخیص حالت چهره چند قومیتی را پیشنهاد کرد. مدل پیشنهادی از سه مرحله تشکیل‌شده است. اولاً مجموعه‌ای از شبکه‌های عصبی باینری آموزش داده می‌شوند، ثانیاً پیش‌بینی‌های این شبکه‌های عصبی برای ایجاد مجموعه ترکیب می‌شوند. و درنهایت از این مجموعه‌ها برای تشخیص وجود یک حالت استفاده می‌شود. پایگاه داده ترکیبی حالت چهره توسط نویسندگان با تصاویری از سه پایگاه داده مختلف ایجادشده است که حاوی تصاویری از موضوعات ژاپنی (JAFFE)، تایوانی (TFEID)، قفقازی (RaFD) و مراکشی است.

نویسندگان نتیجه تشخیص پنج حالت (خشم، شادی، غم، تعجب و ترس) را در دو رویکرد تجربی مختلف گزارش کردند. در مرحله اول، آن‌ها سیستم را در پایگاه داده ترکیبی با دقت ۹۳/۷۵ درصد آموزش و ارزیابی کردند. آزمایش دوم برای ارزیابی روش پیشنهادی در یک محیط کنترل آن کمتر بود، انجام شد. این روش با دو پایگاه داده (TFEID و RaFD) آموزش داده شد و در پایگاه داده JAFFE مورد ارزیابی قرار گرفت و دقت ۴۸/۶۷ درصد به دست آمد.

شان و همکاران [۸] مطالعه‌ای را با استفاده از الگوهای باینری محلی[1](LBP) به‌عنوان استخراج‌کننده انجام داد. آن‌ها تکنیک‌های مختلف یادگیری ماشین مانند تطبیق الگو[2]، ماشین بردار پشتیبانی[3](SVM)، تجزیه‌وتحلیل تشخیص خطی و برنامه‌ریزی خطی را برای تشخیص حالات چهره ترکیب و مقایسه کردند. نویسندگان همچنین مطالعه‌ای را برای تجزیه‌وتحلیل تأثیر وضوح تصویر انجام دادند و به این نتیجه رسیدند که روش‌های مبتنی بر ویژگی‌های هندسی تصاویر با وضوح پایین را به‌خوبی مدیریت نمی‌کنند، درحالی‌که روش‌هایی که بر اساس ظاهر هستند، مانند موجک‌های گابور[4] و LBP، چنین نیستند. بهترین نتیجه به‌دست‌آمده در کار آن‌ها دقت ۹۵/۱ درصد با استفاده از SVM و LBP در پایگاه داده CKþ بود. با استفاده از اعتبارسنجی بین پایگاه داده (آموزش با CKþ و آزمایش با JAFFE) برای ارزیابی سیستم پیشنهادی، نویسندگان به‌دقت ۴۱/۳ درصد دست یافتند. تصاویر ابتدا با استفاده از موقعیت‌های چشمی برش داده شدند. روش مورداستفاده، اعتبارسنجی متقابل ۱۰ برابری بدون همپوشانی موضوعی بود. زمان آموزش و شناسایی توسط نویسندگان ذکر نشده است.

[1] Local binary patterns
[2] Template matching
[3] Support vector machine
[4] Gabor wavelets

یک سیستم تشخیص حالت چهره مبتنی بر ویدئو توسط بایون و کواک [۱۴] پیشنهاد شد. آن‌ها یک D-CNN۳ با توالی تصویر (از خنثی تا حالت نهایی) با استفاده از ۵ فریم متوالی به‌عنوان ورودی ۳ بعدی توسعه دادند. بنابراین، ورودی CNN H W* *5 است (که در آن H و W به ترتیب ارتفاع و عرض تصویر و ۵ تعداد فریم‌ها هستند). نویسندگان ادعا می‌کنند که روش سه‌بعدی CNN می‌تواند درجاتی از تغییر شکل را کنترل کند. با این رویکرد، آن‌ها به‌دقت ۹۵ درصدی دست یافتند، اما این روش متکی به دنباله‌ای است که شامل حرکت کامل از حالت خنثی به حالت‌های دیگر است. آزمایش‌ها با ۱۰ نفر بر روی یک مجموعه داده غیرمعمول انجام شد. زمان آموزش و شناخت توسط نویسندگان ذکر نشده است.

یکی دیگر از رویکردهای مبتنی بر ویدئو، پیشنهادشده توسط فن و تجاهدی [۱۶]، از یک چارچوب مکانی-زمانی[1] مبتنی بر هیستوگرام گرادیان‌ها و جریان نوری استفاده کرد. روش آن‌ها شامل سه مرحله است: پیش‌پردازش، استخراج ویژگی و طبقه‌بندی. در مرحله پیش‌پردازش، تشخیص نشانه‌های چهره انجام شد و یک هم‌ترازی صورت (به‌منظور کاهش تغییرات در وضعیت سر) انجام شد. در مرحله استخراج ویژگی، چارچوبی که اطلاعات پویا استخراج‌شده از تغییر شکل صورت را یکپارچه می‌کند، به کار گرفته شد. در مرحله آخر، طبقه‌بندی، از یک طبقه‌بندی کننده SVM با هسته RBF استفاده شد. آزمایش‌ها با استفاده از پایگاه‌های داده CKþ و MMI انجام شد. دقت به‌دست‌آمده توسط نویسندگان در پایگاه داده CKþ برای هفت عبارت ۸۳/۷ درصد و در پایگاه داده MMI ۷۴/۳ درصد بود. زمان آموزش ذکر نشد، درحالی‌که زمان تشخیص در پایگاه داده CKþ حدود ۳۵۰ میلی ثانیه در هر تصویر و در پایگاه داده MMI، ۵۲۰ میلی ثانیه بود.

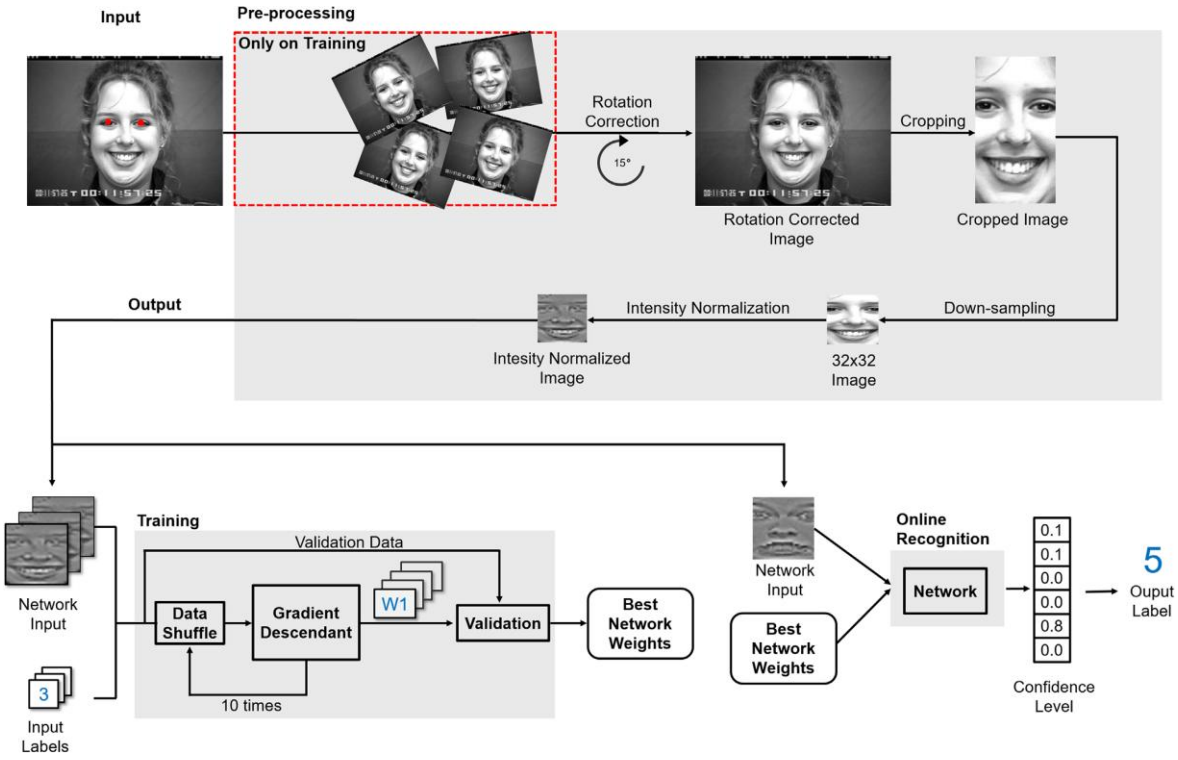
در مقایسه با روش‌های بالا، این مقاله دقت بالاتری در پایگاه‌های داده CKþ و JAFFE (شامل اعتبارسنجی بین پایگاه‌های داده) و زمان آموزش و ارزیابی کمتری نسبت به لئو و همکاران ارائه می‌کند [۱۲،۷]، شان و همکاران [۸] و فن و تجهجادی [۱۶]، یک روش ارزیابی قوی‌تر (بدون همپوشانی موضوعی بین آموزش و آزمون) نسبت به سانگ و همکاران داشتند. بوکرت و همکاران [۱۱] و علای و همکاران [۱۳] روش پیشنهادی را بر روی سه پایگاه داده انجام دادند تا امکان مقایسه منصفانه با سایر روش‌های موجود را به‌جای استفاده از پایگاه‌های داده غیرعمومی مانند Byeon و Kwak فراهم کند [۱۴]. بسیاری از کارهایی که در اینجا ذکر شد دقت بسیار بالایی دارند که نمی‌توان آن را با روش ما مقایسه کرد زیرا امکان همپوشانی موضوعات را در مجموعه‌های آموزشی و آزمایشی فراهم می‌کند. آزمایش‌های اولیه انجام‌شده با روش این مقاله با در نظر گرفتن همپوشانی نیز دقت نزدیک به ۱۰۰ درصد را نشان می‌دهد.

[1] Spatial–temporal

۳– سیستم تشخیص حالت چهره

سیستم ما برای تشخیص حالت چهره سه مرحله یادگیری را فقط در یک طبقه‌بندی (CNN) انجام می‌دهد. سیستم پیشنهادی در دو مرحله اصلی آموزش و آزمایش عمل می‌کند. در طول آموزش، سیستم یک داده آموزشی شامل تصاویر سیاه‌وسفید از چهره‌ها با شناسه حالت مربوطه و مکان‌های مرکز چشم دریافت می‌کند و مجموعه‌ای از وزن‌ها را برای شبکه می‌آموزد. برای اطمینان از اینکه عملکرد آموزش تحت تأثیر ترتیب ارائه نمونه‌ها قرار نگیرد، چند تصویر به‌عنوان اعتبار سنجی از هم جدا می‌شوند و برای انتخاب بهترین مجموعه وزن‌ها استفاده می‌شوند. در طول آزمایش، سیستم تصویری در مقیاس خاکستری از یک صورت به همراه مکان‌های مرکز چشم مربوطه دریافت می‌کند و با استفاده از وزن‌های شبکه نهایی که در طول آموزش آموخته می‌شود، حالت پیش‌بینی‌شده را خروجی می‌دهد.

یک نمای کلی از سیستم در شکل ۲ نشان داده‌شده است. در مرحله آموزش، تصاویر جدید به‌صورت مصنوعی برای افزایش اندازه پایگاه داده تولید می‌شوند. پس‌ازآن، چرخش انجام می‌شود تا چشم‌ها با محور افقی تراز شوند. سپس، تصویر برای حذف اطلاعات پس‌زمینه و حفظ ویژگی‌های خاص برش داده می‌شود.



شکل ۲: نمای کلی سیستم تشخیص حالت چهره پیشنهادی

پس‌ازآن، شدت تصویر نرمال می‌شود. از تصاویر نرمال شده برای آموزش شبکه عصبی کانولوشن استفاده می‌شود. خروجی مرحله آموزش، مجموعه‌ای از وزن‌ها است که با داده‌های اعتبارسنجی پس از

چند دور تمرین بهترین نتیجه را به دست آورد. مرحله آزمایش از روشی مشابه مرحله آموزش استفاده می‌کند. نرمال‌سازی فضایی، برش، نمونه‌برداری پایین و نرمال کردن شدت روشنایی. عبارات به‌صورت اعداد صحیح نشان داده می‌شوند (۰ - عصبانی، ۱ - ناراحتی، ۲ - ترس، ۳ - خوشحال، ۴ - غمگین و ۵ - تعجب).

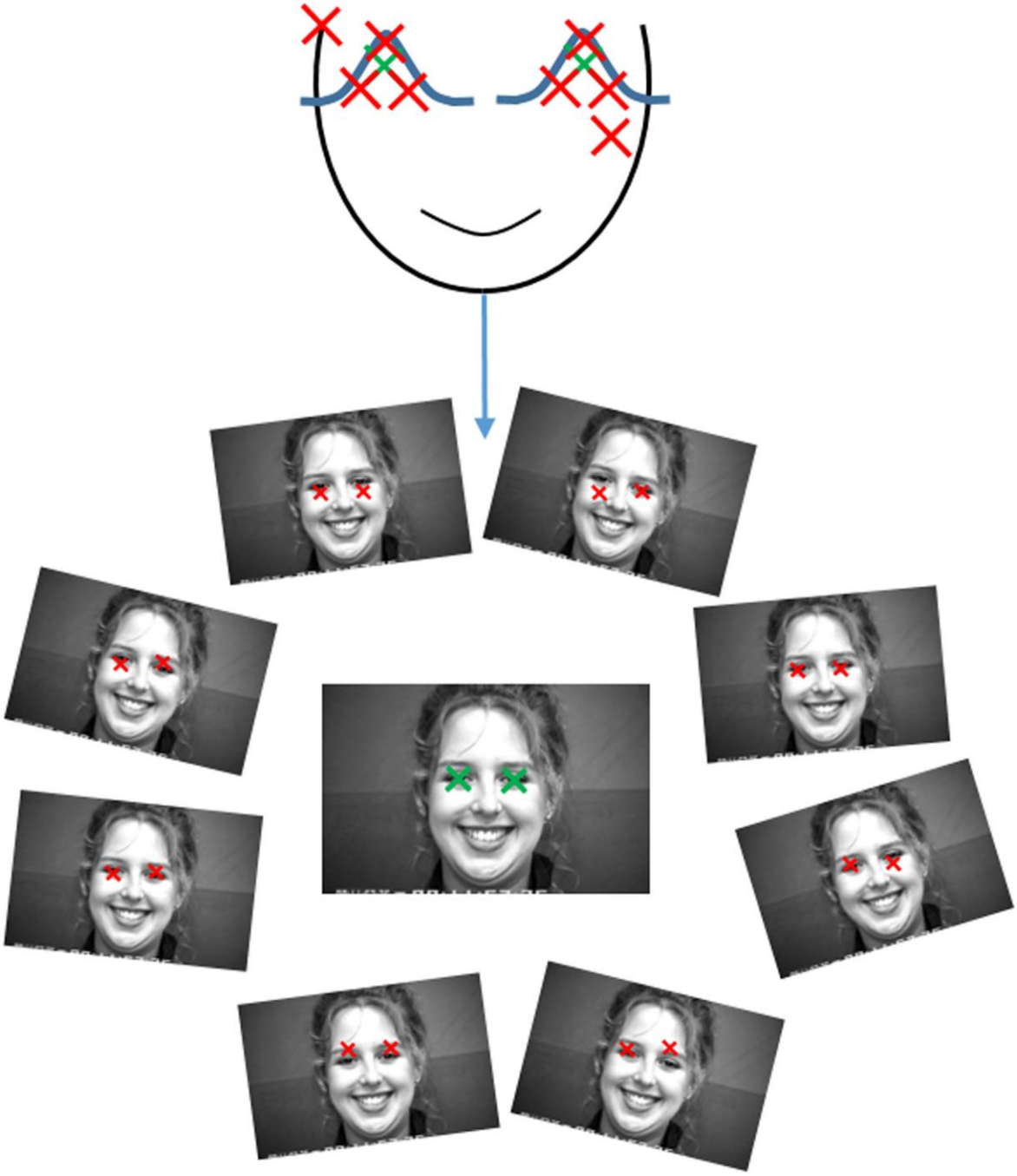
۳-۱- تولید نمونه مصنوعی

نرمال‌سازی فضایی بکار رفته برای اطمینان از اینکه چشم‌ها به‌درستی تراز می‌شوند، کافی نیست. CNN ها در یادگیری توابع تغییرناپذیر بسیار خوب هستند (یعنی می‌توانند تصاویر تحریف‌شده را کنترل کنند [۵۶]). بااین‌حال، یکی از مشکلات اصلی روش‌های یادگیری عمیق این است که برای انجام صحیح این کار به داده‌های زیادی در مرحله آموزش نیاز دارند [۵۶]. متأسفانه، مقدار داده‌های موجود در مجموعه داده‌های عمومی برای دستیابی به آن کافی نیست.

برای رفع این مشکل سیمارد و همکاران [۵۶] تولید تصاویر مصنوعی (یعنی تصاویر واقعی با چرخش مصنوعی، انتقال و انحراف) را برای افزایش پایگاه داده پیشنهاد کرد. این فرآیند به‌عنوان تقویت داده[1] نامیده می‌شود. نویسندگان مزایای استفاده از ترکیبی از انتقال، چرخش و انحراف را برای افزایش پایگاه داده نشان می‌دهند. به دنبال این ایده، در این مقاله، از توزیع گاوسی دوبعدی ($\sigma = 3$ و $\mu = 0$) برای معرفی نویز تصادفی در مکان‌های مرکز چشم استفاده می‌کنیم. برای هر تصویر، ۷۰ تصویر مصنوعی اضافی تولید شد.

همان‌طور که در شکل ۳ مشاهده می‌شود، نقاط برای هر دو چشم به دنبال توزیع گاوسی در مرکز اصلی آن‌ها ایجاد می‌شود. بنابراین موقعیت جدید مرکز چشم معادل موقعیت اصلی است، اما با نویز گاوسی مختل می‌شود. ازآنجایی‌که مقادیر جدید داده‌شده به روش نرمال‌سازی مرکز چشم نیستند، تصاویر به‌دست‌آمده یا توسط انتقال، چرخش و/یا مقیاس مختل می‌شوند یا اصلاً مختل نمی‌شوند. مقدار خاص انحراف استاندارد گاوسی باید با دقت انتخاب شود. یک انحراف بسیار کوچک می‌تواند تغییری در داده‌های اصلی ایجاد نکند و تصاویر شبیه به هم و بی‌فایده ایجاد کند. از سوی دیگر، یک انحراف بزرگ برای هر چشم می‌تواند باعث ایجاد انتقال، چرخش و/یا نویز بیش‌ازحد در تصاویر شود و طبقه‌بندی کننده برای یادگیری ویژگی‌های حالت پیچیده‌تر می‌شود. انحراف استاندارد $\sigma = 3$ به‌صورت تجربی انتخاب شد. توجه به این نکته ضروری است که داده‌های مصنوعی فقط در آموزش استفاده می‌شود.

[1] Data augmentation



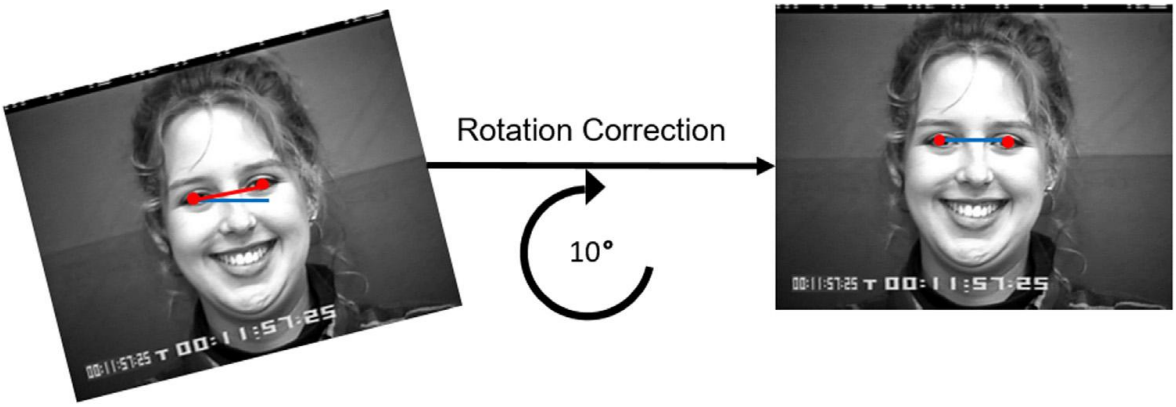
شکل ۳: تصویری از تولید نمونه مصنوعی

۳-۲- تصحیح چرخش

تصاویر موجود در پایگاه‌های داده و همچنین در محیط‌های واقعی، ازنظر چرخش، روشنایی و اندازه حتی برای تصاویر یک سوژه متفاوت هستند. این تغییرات به حالت چهره مربوط نمی‌شود و می‌تواند بر میزان دقت سیستم تأثیر بگذارد. برای رفع این مشکل، ناحیه صورت (با نرمال‌سازی چرخش) با افق و یک نقطه مرکزی تراز می‌شود تا مسائل هندسی احتمالی مانند چرخش‌ها و انتقال‌ها را اصلاح کند. برای انجام این هم‌ترازی به دو اطلاعات، تصویر صورت و مرکز هر دو چشم نیاز است. در حال حاضر روش‌های

زیادی وجود دارد که می‌توانند چشم‌ها و سایر نقاط صورت را با دقت بالا بیابند [۵۷-۶۱]، و این موضوع کار این مقاله نیست. چنگ و همکاران [۶۱] یک نسخه CUDA از DRMF [۶۰] را توسعه داد، که امکان تشخیص لحظه‌ای نقاط کلیدی صورت را فراهم می‌کرد.

برای انجام تراز چهره، یک تبدیل چرخشی برای تراز کردن چشم‌ها با محور افقی تصویر و یک انتقال پیوندی که توسط مکان‌های چشم‌ها تعریف‌شده است برای متمرکز کردن صورت در یک نقطه خاص از تصویر اعمال می‌شود. چرخش باعث می‌شود که زاویه تشکیل‌شده توسط پاره‌خط از یک مرکز چشم به مرکز دیگر و محور افقی صفر شود. چرخش‌ها و انتقال‌ها در تصاویر مربوط به حالت چهره نیست و بنابراین باید حذف شوند تا بر میزان دقت سیستم تأثیر منفی نگذارند. روش اصلاح چرخش در شکل ۴ نشان داده‌شده است.



شکل ۴: مثال تصحیح چرخش

ورودی این روش می‌تواند یک تصویر اصلی یا مصنوعی باشد. تصحیح چرخش، برای تصاویر تولیدشده مصنوعی، ممکن است تراز کاملی با محور افقی نداشته باشد زیرا مرکز چشم موقعیت واقعی است که توسط یک نویز تصادفی گاوسی مختل شده است. بنابراین، تصاویری ایجاد می‌کند که توسط چرخش‌ها و ترجمه‌ها مختل می‌شوند، که تنوع در نمونه‌های آموزشی را افزایش می‌دهد.

۳-۳- برش تصویر

همان‌طور که در شکل ۲ نشان داده‌شده است، تصویر اصلی دارای اطلاعات پس‌زمینه زیادی است که برای روش طبقه‌بندی مهم نیست. این اطلاعات می‌تواند دقت طبقه‌بندی را کاهش دهد زیرا طبقه‌بندی کننده یک مشکل دیگر برای حل دارد، یعنی تمایز بین پس‌زمینه و پیش‌زمینه. پس از برش، تمام

قسمت‌های تصویری که بیان خاصی در شکل‌گیری ندارند حذف می‌شوند. ناحیه کاشت همچنین سعی می‌کند قسمت‌هایی از صورت را که برای بیان نقشی ندارند (مانند گوش‌ها، قسمتی از پیشانی و غیره) حذف کند. بنابراین، منطقه موردنظر بر اساس نسبت فاصله بین چشم‌ها تعریف می‌شود. درنتیجه، روش ما قادر است افراد و اندازه‌های مختلف تصویر را بدون دخالت انسان مدیریت کند. ناحیه کاشت با ضریب عمودی ۵/۴ (با در نظر گرفتن ۳/۱ برای ناحیه بالای چشم و ۲/۳ برای ناحیه زیر چشم) تعیین می‌شود که به فاصله بین نقطه وسط چشم و مرکز چشم راست اعمال می‌شود. منطقه برداشت افقی با ضریب ۴/۲ که در همین فاصله اعمال می‌شود محدود می‌شود. این مقادیر به‌صورت تجربی تعیین شد. نمونه‌ای ازاین‌رو در شکل ۵ نشان داده‌شده است.

شکل ۵: نمونه برش تصویر

۳–۴– نمونه‌برداری پایین

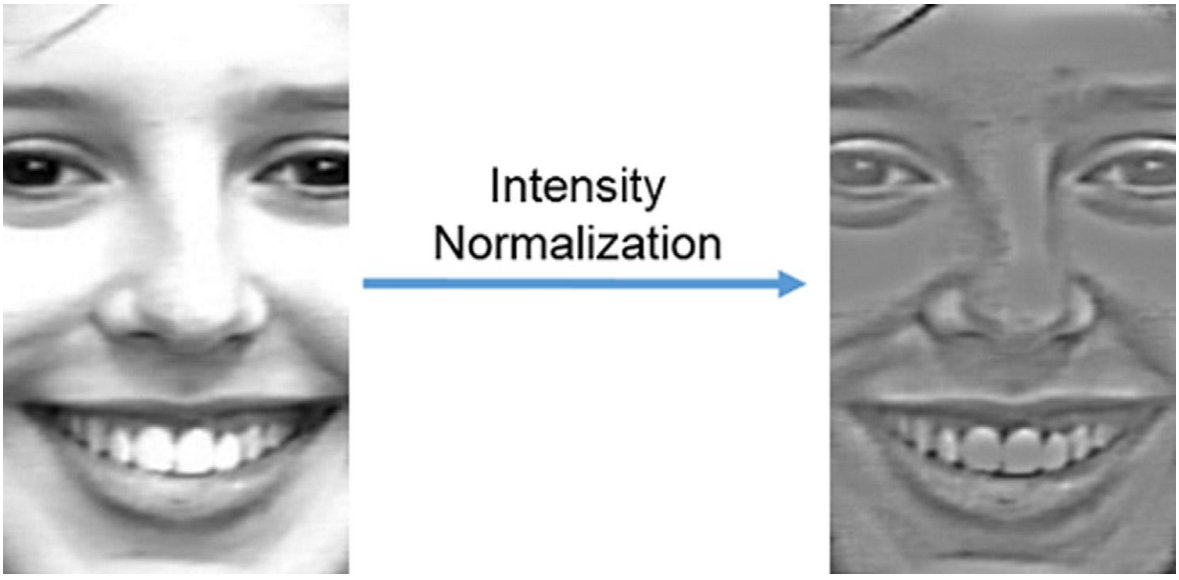
عملیات نمونه‌برداری برای کاهش اندازه تصویر برای شبکه و اطمینان از عادی‌سازی مقیاس انجام می‌شود. مکان یکسان برای اجزای صورت (چشم، دهان، ابرو و غیره) در همه تصاویر. نمونه‌گیری پایین از یک رویکرد درون‌یابی خطی استفاده می‌کند. پس‌ازاین نمونه‌برداری مجدد، می‌توان تضمین کرد که مرکز چشم تقریباً در همان موقعیت قرار می‌گیرد. این روش به CNN کمک می‌کند تا یاد بگیرد که کدام مناطق به هر عبارت خاص مرتبط هستند. نمونه‌برداری پایین همچنین امکان انجام کانولوشن در GPU را فراهم می‌کند زیرا امروزه اکثر کارت‌های گرافیک حافظه محدودی دارند. تصویر نهایی با استفاده از درون‌یابی خطی به ۳۲ × ۳۲ پیکسل نمونه‌برداری می‌شود.

۳–۵– نرمال‌سازی شدت روشنایی

روشنایی و کنتراست تصویر می‌تواند حتی در تصاویر یک شخص در یک حالت متفاوت باشد. بنابراین، تغییر در بردار ویژگی افزایش می‌یابد. چنین تغییراتی پیچیدگی مسئله‌ای را که طبقه‌بندی کننده باید

برای هر حالت حل کند، افزایش می‌دهد. به‌منظور کاهش این مسائل یک نرمال‌سازی شدت روشنایی اعمال شد. یک روش [۶۲] به نام نرمال‌سازی کنتراست[1]، استفاده شد. اساساً نرمال‌سازی دومرحله‌ای است. ابتدا یک نرمال‌سازی کنتراست موضعی کاهشی انجام می‌شود. و در مرحله دوم، یک نرمال‌سازی کنتراست محلی تقسیمی اعمال می‌شود. در مرحله اول، مقدار هر پیکسل از میانگین وزنی گاوسی همسایه‌های آن کم می‌شود. در مرحله دوم، هر پیکسل بر انحراف معیار همسایگی خود تقسیم می‌شود. همسایگی برای هر دو مرحله از یک هسته ۷×۷ پیکسل (به‌صورت تجربی انتخاب‌شده) استفاده می‌کند. نمونه‌ای ازاین‌رو در شکل ۶ نشان داده‌شده است. معادله ۱ نشان می‌دهد که چگونه هر مقدار پیکسل جدید درروش عادی‌سازی شدت محاسبه می‌شود:

$$x' = \frac{x - \mu_{nhgx}}{\sigma_{nhgx}} \qquad (۱)$$



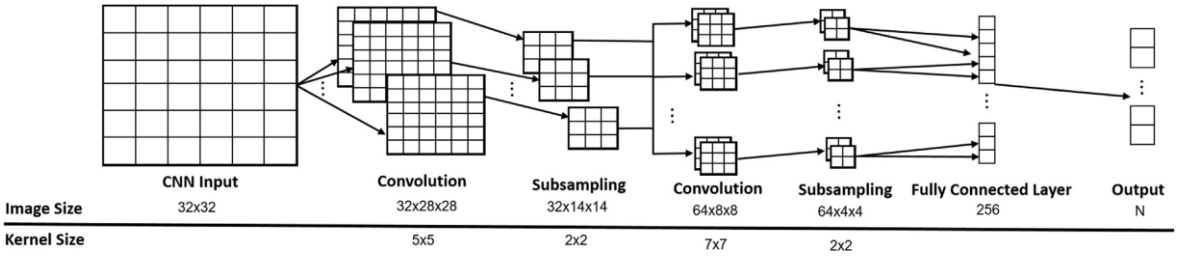
شکل ۶: تصویر نرمال‌سازی شدت روشنایی

۳-۶- شبکه عصبی کانولوشن

معماری شبکه عصبی کانولوشن ما در شکل ۷ ارائه‌شده است. شبکه یک تصویر ۳۲×۳۲ در مقیاس خاکستری را به‌عنوان ورودی دریافت می‌کند. معماری CNN شامل ۲ لایه کانولوشن، ۲ لایه نمونه‌برداری فرعی و یک‌لایه کاملاً متصل است. اولین لایه CNN یک‌لایه کانولوشن است که یک هسته کانولوشن ۵×۵ را اعمال می‌کند و ۳۲ تصویر ۲۸×۲۸ پیکسل را خروجی می‌دهد. این لایه توسط یک‌لایه نمونه‌گیری

[1] Contrastive equalization

فرعی دنبال می‌شود که از max-pooling (با اندازه هسته ۲×۲) استفاده می‌کند تا تصویر را به نصف اندازه آن کاهش دهد. پس‌ازآن، یک‌لایه کانولوشن جدید ۶۴ لایه را با یک هسته ۷ × ۷ برای لایه قبلی انجام می‌دهد و پس‌ازآن یک نمونه فرعی دیگر، دوباره با یک هسته ۲×۲ دنبال می‌شود. خروجی‌ها به یک‌لایه پنهان کاملاً متصل داده می‌شود که دارای ۲۵۶ نورون است. درنهایت، شبکه دارای شش یا هفت گره خروجی است که به‌طور کامل به لایه قبلی متصل هستند. اولین لایه شبکه (لایه پیچیدگی) باهدف استخراج ویژگی‌های بصری اولیه، مانند لبه‌های جهت‌دار، نقطه پایانی، گوشه‌ها، توسط Lecun و همکارانش توضیح داده‌شده است [۳۶].



شکل ۷: معماری شبکه کانولوشن پیشنهادی

در مشکل تشخیص حالت چهره، ویژگی‌های شناسایی‌شده عمدتاً شکل، گوشه و لبه چشم، ابرو و لب است. هنگامی‌که ویژگی‌ها شناسایی می‌شوند، مکان دقیق آن چندان مهم نیست، فقط موقعیت نسبی آن در مقایسه با سایر ویژگی‌ها مهم است. به‌عنوان‌مثال، موقعیت مطلق ابروها مهم نیست، اما فاصله آن‌ها از چشم مهم است، زیرا فاصله زیاد ممکن است به‌عنوان‌مثال بیانگر تعجب باشد. لایه دوم (یک‌لایه نمونه فرعی) وضوح فضایی نقشه ویژگی را کاهش می‌دهد. در مطالعه Lecun و همکاران [۳۶]، هدف عملیات کاهش دقت کدگذاری موقعیت ویژگی‌های استخراج‌شده توسط لایه قبلی است. دولایه بعدی، یکی کانولوشنال و نمونه فرعی، باهدف انجام عملیات مشابه با لایه‌های اول، اما مدیریت ویژگی‌ها در سطح پایین‌تر، شناسایی عناصر متنی (عناصر چهره) به‌جای اشکال، لبه‌ها و گوشه‌های ساده است. آخرین لایه پنهان (لایه کاملاً متصل) مجموعه‌ای از ویژگی‌های آموخته‌شده را دریافت می‌کند و سطح اطمینان ویژگی‌های داده‌شده را در هر یک از عبارات در نظر گرفته‌شده خروجی می‌دهد.

این شبکه از روش شیب نزولی تصادفی برای محاسبه وزن‌های سیناپسی بین نورون‌ها استفاده می‌کند، این روش توسط Buttou [۶۳] ارائه‌شده است. ارزش اولیه این سیناپس‌ها برای پیچش‌ها و برای لایه کاملاً متصل با استفاده از پرکننده Xavier، پیشنهادشده توسط Glorot و همکاران، ایجاد می‌شود [۶۴]، که به‌طور خودکار مقیاس اولیه را بر اساس تعداد نورون‌های ورودی و خروجی تعیین می‌کند. تلفات با استفاده از یک تابع لجستیک از خروجی Soft-max (معروف به Soft maxWithLoss) محاسبه

می‌شود. تابع فعال‌سازی نورون‌ها یک ReLu (واحد خطی اصلاح‌شده) است که به‌صورت f(z)=max(z,0) تعریف می‌شود. تابع ReLu به‌طورکلی در معماری‌های عمیق بسیار سریع‌تر یاد می‌گیرد [۶۵].

۴– آزمایش‌ها و بحث‌ها

آزمایش‌ها با استفاده از سه پایگاه داده در دسترس عموم درزمینه تحقیقاتی تشخیص حالات صورت انجام شد. پایگاه داده Cohn-Kanade (CKþ)، پایگاه داده بیان‌های صورت زنانه ژاپنی JAFFE و بیان‌های چهره سه‌بعدی دانشگاه Binghamton. پایگاه داده BU-3DFE. با در نظر گرفتن یک طبقه‌بندی کننده برای طبقه‌بندی تمام حالات آموخته‌شده دقت محاسبه می‌شود. علاوه بر این، برای امکان مقایسه بهتر با برخی از روش‌ها، دقت نیز با در نظر گرفتن یک طبقه‌بندی باینری برای هر عبارت، همان‌طور که در [۷] استفاده‌شده است، محاسبه می‌شود.

اجرای مراحل پیش‌پردازش در داخل با استفاده از OpenCV، Cþ þ و یک کتابخانه CNN مبتنی بر GPU انجام شد. تمام آزمایش‌ها با استفاده از یک Intel Core i7 3.4 گیگاهرتز با NVIDA GeForce GTX 660 CUDA Capable که دارای ۵/۱ گیگابایت حافظه در GPU است انجام شد. محیط آزمایش‌ها لینوکس اوبونتو ۰۴/۱۲، با NVIDIA CUDA Framework 6.5 و کتابخانه cuDNN نصب‌شده بود. مرحله پیش‌پردازش (تصحیح چرخش، برش، نمونه‌برداری پایین و نرمال‌سازی شدت روشنایی) تنها ۰/۲۰ ثانیه و تشخیص شبکه (مرحله طبقه‌بندی) به‌طور متوسط ۰/۰۱ ثانیه طول کشید.

در این بخش، مطالعه‌ای انجام‌شده است که تأثیر هر مرحله نرمال‌سازی را در دقت روش نشان می‌دهد. در ابتدا، ما پایگاه‌های داده مورداستفاده برای آزمایش‌ها را توصیف می‌کنیم. در مرحله دوم، معیارهای مورداستفاده برای ارزیابی دقت سیستم توضیح داده‌شده است. سوم، نتایج آزمایش‌های تأثیر هر مرحله پیش‌پردازش ارائه‌شده است. چهارم، نتایج با پایگاه‌های اطلاعاتی مختلف نشان داده‌شده و به‌تفصیل موردبحث قرارگرفته است. درنهایت، مقایسه‌ای با چندین روش اخیر تشخیص چهره که از روش ارزیابی یکسانی استفاده می‌کند، ارائه‌شده است و محدودیت‌های روش ما موردبحث قرار می‌گیرد.

۴–۱– پایگاه داده

سیستم ارائه‌شده با استفاده از پایگاه داده CKþ، پایگاه داده JAFFE و پایگاه داده BU-3DFE آموزش و آزمایش شد. پایگاه داده CKþ شامل ۱۰۰ دانشجو با سن بین ۱۸ تا ۳۰ سال است. افراد در پایگاه داده ۶۵ درصد زن، ۱۵ درصد آفریقایی-آمریکایی و ۳ درصد آسیایی یا آمریکای جنوبی هستند. این تصاویر از دوربینی که دقیقاً در مقابل سوژه قرار دارد گرفته‌شده است. به دانش آموزان دستور داده شد که یک سری حالات را اجرا کنند. تمام تصاویر موجود در پایگاه داده دارای آرایه‌های ۶۴۰ در ۴۸۰ پیکسل با دقت ۸ بیتی برای مقادیر خاکستری هستند. هر تصویر دارای یک فایل توصیف‌کننده با نقاط صورت خود

است، از این نقاط برای نرمال‌سازی تصویر حالت چهره استفاده‌شده است. نقاط چهره در پایگاه داده با استفاده از سیستم کدگذاری کنش صورت (FACS) کدگذاری می‌شوند [۶۶]. سازندگان پایگاه داده از مدل‌های ظاهری فعال (AAMs) برای استخراج خودکار نقاط صورت استفاده کردند. پایگاه داده حاوی تصاویری برای عبارات خنثی، عصبانی، تحقیر، انزجار، ترس، خوشحالی، غم و تعجب است. برای مقایسه منصفانه با بخش عمده روش‌های اخیر [۸،۶۷،۴۳،۴۴،۷،۱۶]، در آزمایش‌های ما از تصاویر حالت تحقیر استفاده نشد. چند نمونه از تصاویر پایگاه داده CKþ در شکل ۸ نشان داده‌شده است.

شکل ۸: نمونه‌ای از تصاویر در پایگاه داده CKþ

برای انجام یک ارزیابی درست از روش پیشنهادی، پایگاه داده در ۸ گروه بدون همپوشانی موضوعی بین گروه‌ها جدا شد (یعنی اگر تصویر یک موضوع در یک گروه باشد، هیچ تصویری از همان موضوع در هیچ گروه دیگری وجود نخواهد داشت). هر گروه شامل ۱۲ موضوع است. این روش تضمین می‌کند که گروه‌های آزمایشی موضوعاتی از گروه آموزشی ندارند و همچنین با روش‌های بسیاری که قبلاً معرفی‌شده استفاده می‌شود [۶۷،۴۳،۴۴،۱۶-۷۰،۷،۸،۶۹]. همان‌طور که توسط Girard و همکاران بحث شده است [۵۳]، این روش (موضوعات مختلف در گروه‌های آموزشی/آزمایی و اعتبارسنجی متقابل) قابلیت تعمیم طبقه‌بندی کننده‌ها را تضمین می‌کند. زاواشی و همکاران [۴۴] نیز در این موردبحث و گفتگو کردند. در یک آزمایش، گروه‌ها حاوی تصاویری از یک موضوع (نه تصاویر مشابه) هستند، درحالی‌که در آزمایش دیگر تضمین می‌کنند که تصاویر همان موضوع به‌طور هم‌زمان در گروه‌های آموزشی و آزمایشی قرار ندارند. که در آزمایش اول دقت ۹۹/۴۰ درصد به دست آمد، درحالی‌که در آزمایش دوم دقت به ۹۰/۸۸ درصد کاهش می‌یابد. این نتیجه نشان می‌دهد که روش‌هایی که بدون موضوعات مشابه در گروه‌های آموزش/آزمون ارزیابی می‌شوند (که ما معتقدیم ارزشیابی منصفانه‌تری است) عموماً دقت پایین‌تری نسبت به آن‌هایی که این محدودیت را تضمین نمی‌کنند ارائه می‌دهند. ما همچنین این نتایج را با آزمایش‌های اولیه با استفاده از روش خود تأیید کردیم.

برای تأیید روش پیشنهادی، آزمایش‌های متقابل پایگاه داده نیز انجام شد. این آزمایش‌ها از پایگاه‌های داده JAFFE و BU-3DFE استفاده کردند. پایگاه داده JAFFE از ۲۱۳ تصویر از ۱۰ سوژه زن ژاپنی تشکیل‌شده است. در این پایگاه داده در هر یک از شش حالت اصلی حدود ۴ تصویر و از هر موضوع یک

تصویر از عبارت خنثی وجود دارد. تمام تصاویر موجود در مجموعه داده، آرایه‌های پیکسلی ۲۵۶ در ۲۵۶ با دقت ۸ بیتی برای مقادیر مقیاس خاکستری هستند. ازآنجایی‌که این پایگاه داده کوچک‌تر است، گروه‌ها بر اساس موضوع از هم جدا می‌شوند، یعنی هر گروه حاوی تصاویری از یک موضوع است، بنابراین ۱۰ گروه تشکیل شد. چند نمونه از تصاویر پایگاه داده JAFFE در شکل ۹ نشان داده‌شده است.

شکل ۹: نمونه‌ای از تصاویر موجود در پایگاه داده JAFFE

پایگاه داده دیگری که برای ارزیابی روش پیشنهادی استفاده می‌شود، پایگاه داده دانشگاه بینگهامتون (BU-3DFE) است [۶]. پایگاه داده BU-3DFE شامل ۶۴ نفر (۵۶ درصد زن و ۴۴ درصد مرد) است که در محدوده سنی بین ۱۸ سال تا ۷۰ سال قرار دارند، با انواع اجداد قومی/نژادی، ازجمله سفید، سیاه‌پوست، آسیای شرقی، آسیای میانه، هندی. و اسپانیایی لاتین. تمام تصاویر موجود در مجموعه داده آرایه‌های پیکسلی ۱۵۶ در ۲۰۹ هستند. این پایگاه نیز در ۸ گروه بدون همپوشانی موضوعی بین گروه‌ها تفکیک شد. هر گروه شامل حدود ۸ موضوع است. چند نمونه از تصاویر پایگاه داده BU-3DFE در شکل ۱۰ نشان داده‌شده است.

شکل ۱۰: نمونه‌ای از تصاویر در پایگاه داده BU-3DFE

همه پایگاه‌های داده حاوی تعداد زیادی تصاویر از موضوعات مشابه برای هر عبارت هستند و این تصاویر بسیار شبیه به یکدیگر هستند. بنابراین، از هر موضوع، تنها ۳ فریم از هر عبارت (یعنی رساترین

قاب‌ها) و ۱ فریم برای بیان خنثی (یعنی کم رساترین قاب) از هر موضوع استفاده‌شده است. با پیروی ازاین‌رو، اندازه‌های پایگاه داده به‌دست‌آمده به این شرح است، برای پایگاه داده CKþ 2100 نمونه (بدون نمونه‌های مصنوعی) و ۱۴۷۰۰۰ نمونه (با نمونه‌های مصنوعی)، برای پایگاه داده JAFFE 213 نمونه (بدون نمونه‌های مصنوعی) و ۱۴۹۱۰ نمونه (بدون نمونه‌های مصنوعی) و برای پایگاه داده BU-3DFE 1344 نمونه (بدون نمونه‌های مصنوعی) و ۹۴۰۸۰ نمونه (با نمونه‌های مصنوعی).

۴-۲- معیارهای ارزیابی

برای مقایسه منصفانه روش ارائه‌شده با سایر روش‌ها، دقت به دو روش مختلف محاسبه شد. در حالت اول، یک طبقه‌بندی کننده برای تمام حالات پایه استفاده می‌شود. دقت به‌سادگی با استفاده از میانگین، Cnclass، دقت طبقه‌بندی کلاس n در هر عبارت، CnclassE، یعنی تعداد بازدیدهای یک عبارت به ازای مقدار داده آن عبارت، محاسبه می‌شود. (معادلات ۲)

$$c_{nclass} = \frac{\sum c_{nclassE}}{n}, \qquad c_{nclassE} = \frac{Hit_E}{T_E} \tag{۲}$$

که در آن Hit تعداد بازدیدها در حالت E است، T تعداد کل نمونه‌های آن حالت و n تعداد حالاتی است که باید در نظر گرفته شوند.

در مرحله دوم، یک طبقه‌بندی باینری برای هر عبارت، طبقه‌بندی یک در مقابل همه را انجام می‌دهد. با استفاده ازاین‌رویکرد، تصاویر به n طبقه‌بندی باینری ارائه می‌شوند که n تعداد حالاتی است که طبقه‌بندی می‌شوند. هدف هر طبقه‌بندی کننده این است که اگر تصویر حاوی یک عبارت خاص باشد، بله یا در غیر این صورت خیر. به‌عنوان‌مثال، اگر یک تصویر حاوی عبارت شگفت‌انگیز باشد، طبقه‌بندی شگفت‌انگیز باید بله و تمام پنج طبقه‌بندی دیگر باید نه پاسخ دهند. تنها تفاوت این طبقه‌بندی کننده با معماری ارائه‌شده در بخش ۳ این است که برای هر طبقه‌بندی کننده تنها دو خروجی لازم است(معادله۳).

$$c_{bin} = \frac{\sum c_{binE}}{n} \; c_{binE} = \frac{Hit_E + Hit_{NE}}{T} \tag{۳}$$

۴-۳- تنظیم پیش‌پردازش

همان‌طور که قبلاً توضیح داده شد، روش پیشنهادی یک مرحله پیش‌پردازش را ترکیب می‌کند که هدف آن حذف ویژگی‌های خاص یک تصویر چهره و یک شبکه عصبی کانولوشنال است.

در این بخش، تأثیر دقت در طبقه‌بندی هر عملیات در مرحله پیش‌پردازش را ارائه می‌کنیم. در اینجا، ما به‌طور تصادفی ترتیب نمونه‌ها را در شبکه تولید می‌کنیم و از یک اعتبارسنجی متقابل k-fold ساده

بین ۸ گروه پایگاه داده CKþ استفاده می‌کنیم. پایگاه داده به دو مجموعه آموزشی (با ۷ گروه) و آزمون (با ۱ گروه) تقسیم شد. این آموزش ۸ بار با استفاده از ۲۰۰۰ دوره برای هر یک از آن‌ها انجام شد. دقت در هر عبارت (C6classE) و میانگین کلی برای همه عبارات (C6class) محاسبه شد.

ابتدا بدون پیش‌پردازش انجام شد. این آزمایش اول با استفاده از پایگاه داده اصلی، بدون هیچ‌گونه مداخله یا پیش‌پردازش تصویر، فقط یک نمونه‌برداری از تصویر به‌اندازه ورودی CNN انجام شد. در این آزمایش، میانگین دقت برای تمام عبارات C6class = 53.57٪ بود. دقت در هر حالت در جدول ۱ نشان داده‌شده است. سپس برای همه اجراها انجام شد. همان‌طور که در جدول ۱ مشاهده می‌شود، تنها با استفاده از CNN بدون هیچ‌گونه پیش‌پردازش تصویر، نرخ تشخیص در مقایسه با روش‌های موجود پایین است. ما معتقدیم که تنوع و تعداد نمونه‌ها در پایگاه داده CKþ اندک بود که به شبکه عصبی کانولوشن اجازه نمی‌داد نحوه برخورد با واریانس، محیط و موضوع را بیاموزد. علاوه بر این، از طیف کامل فضای تصویر موجود در ورودی شبکه برای نمایش چهره استفاده نمی‌کند.

جدول ۱ :مراحل پیش‌پردازش تنظیم برای پایگاه داده CKþ

Preprocessing Step	Angry (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)	Average (%)
A	28.10	51.23	17.91	70.68	20.99	77.52	53.57
B	68.20	79.01	23.37	86.39	23.46	87.16	71.67
C	17.17	79.09	00.00	48.92	05.05	91.25	61.55
D	81.82	90.74	73.13	95.81	66.67	94.50	87.86
E	27.27	52.94	08.22	79.10	18.29	85.54	57.00
F	78.51	93.21	53.73	95.29	75.31	93.12	86.67
G	86.05	88.30	69.33	96.60	77.11	95.34	87.10
H	79.34	94.44	73.13	99.48	72.84	94.94	89.76

قسمت دوم برش تصویر است. به‌منظور افزایش عملکرد روش، همان‌طور که در بخش ۳ توضیح داده شد، تصویر به‌طور خودکار برش داده می‌شود تا مناطق خاصی که مفید نیستند، در هر دو مرحله آموزش و آزمایش حذف شوند. درنتیجه، میانگین دقت برای تمام عبارات به C6class = 71.67٪ افزایش یافت. دقت در هر عبارت در جدول ۱ نشان داده‌شده است. در اینجا، نمونه‌برداری پایین نیز انجام می‌شود، زیرا ورودی شبکه پیشنهادی یک تصویر ثابت ۳۲×۳۲ پیکسل است.

در مقایسه با نتیجه نشان داده‌شده در قبل، می‌توانیم افزایش قابل‌توجهی در نرخ شناسایی را تنها با افزودن فرآیند برش مشاهده کنیم. دلیل اصلی افزایش دقت این است که با برش، بسیاری از اطلاعات غیرضروری را حذف می‌کنیم که طبقه‌بندی کننده برای تعیین بیان موضوع و استفاده بهتر از فضای تصویر موجود در ورودی شبکه به آن‌ها نیاز دارد.

قسمت سوم تصحیح چرخش است. یک تصحیح چرخش (و نمونه‌برداری پایین) در تصویر انجام می‌شود تا چرخش‌هایی را که مربوط به تغییرات چهره (که می‌تواند مختص حالت یا حرکت دوربین باشد) حذف کند، در هر دو مرحله آموزش و آزمایش این کار انجام می‌شود. میانگین دقت برای همه عبارات C6class = 61.55٪ بود.

باید توجه کرد که این نتیجه فقط اصلاح چرخش را اعمال می‌کند. در مقایسه با نتیجه عدم پیش‌پردازش، می‌توان افزایش دقت را در حدود ۸/۰۰ درصد مشاهده کرد. این افزایش ممکن است ناشی از تغییرات کمتری باشد که شبکه باید مدیریت کند. با اصلاح چرخش، عناصر صورت (چشم‌ها، دهان و ابروها) بیشتر در همان فضای پیکسل باقی می‌مانند، اما همچنان تأثیر پس‌زمینه را دارند و از طیف کامل فضای تصویر موجود در ورودی شبکه برای نمایش استفاده نمی‌کنند.

قسمت چهارم عادی‌سازی فضایی است. همان‌طور که قبلاً مشاهده شد، برش تصویر و اصلاح چرخش به‌طور جداگانه اعمال می‌شود و دقت طبقه‌بندی کننده را افزایش می‌دهد. به این دلیل اتفاق می افتد که هر دو روش پیچیدگی را کاهش می‌دهند. در اینجا، ما نرمال‌سازی فضایی کامل را موردبحث قرار می‌دهیم که توسط برش تصویر، تصحیح چرخش و نمونه‌برداری پایین تشکیل‌شده است. با ترکیب عملیات، میانگین دقت برای همه عبارات C6class = 87.86% بود. همان‌طور که انتظار می‌رفت، پیوستن هر دو روش در مرحله پیش‌پردازش باعث افزایش دقت می‌شود. به این دلیل اتفاق می‌افتد که بسیاری از تغییرات غیر مرتبط از تصویر حذف‌شده است. اگرچه شبکه عصبی کانولوشن می‌تواند این تغییرات را مدیریت کند، اما ما به یک پایگاه داده بزرگ‌تر (که نداریم) و شاید معماری پیچیده‌تری نیاز داریم.

قسمت پنجم نرمال‌سازی شدت روشنایی است. روش عادی‌سازی فضایی به‌طور قابل‌توجهی دقت کلی سیستم را افزایش می‌دهد. نرمال‌سازی شدت روشنایی برای حذف تغییرات روشنایی در رنگ استفاده می‌شود. این آزمایش فقط با استفاده از نرمال‌سازی شدت روشنایی انجام شد. از همان روشی که قبلاً توضیح داده شد استفاده می‌کند. میانگین دقت برای همه عبارات C6class = 57.00٪ بود. همان‌طور که مشاهده می‌شود، تنها با اعمال نرمال‌سازی شدت روشنایی، دقت طبقه‌بندی کننده نیز اندکی افزایش یافت.

مرحله ششم نرمال‌سازی فضایی و شدت روشنایی است. با ترکیب نرمال‌سازی‌های فضایی (تصحیح چرخش، برش و نمونه‌برداری پایین) و شدت روشنایی، بخش بزرگی از تغییرات غیر مرتبط باحالت چهره را حذف می‌کنیم و فقط تغییرات خاص را که به ژست یا محیط مربوط نمی‌شود باقی می‌گذاریم. میانگین دقت برای همه حالات C6class = 86.67٪ بود. دقت هر عبارت در جدول ۱ نشان داده‌شده است.

۴-۴- نتایج

همان‌طور که قبلاً بحث شد، یک روش آموزش/آزمایش ساده برای ارزیابی تأثیر مراحل پیش‌پردازش استفاده شد. برخلاف آزمایش‌های تنظیم که فقط از مجموعه‌های آموزشی و آزمایشی استفاده می‌کردند، این بخش پایگاه‌های اطلاعاتی هر آزمایش را در سه مجموعه اصلی انجام می‌دهد. مجموعه آموزشی، مجموعه اعتبارسنجی و مجموعه تست.

همچنین ذکر شد که برای آموزش شبکه از روش نزول گرادیان استفاده‌شده است. چنین روش‌هایی ممکن است تحت تأثیر ترتیب ارائه نمونه‌ها به شبکه باشد که باعث تغییر در دقت می‌شود. همان‌طور که در جدول ۲ مشاهده می‌شود، دقت در حدود ۴/۰۰ درصد افزایش می‌یابد (یا کاهش می‌یابد) که با تغییر ترتیب نمونه‌های آموزشی رخ می‌دهد. جدول ۲ دقت یک آموزش را نشان می‌دهد که ۱۰ بار انجام‌شده است که هرکدام با ترتیب تصادفی نمونه‌های آموزشی ارائه‌شده است. برای اینکه کمتر تحت تأثیر این تغییرات دقت قرار بگیریم، یک روش آموزشی پیشنهاد می‌کنیم که از مجموعه اعتبارسنجی برای انتخاب بهترین وزن شبکه بر اساس آموزش‌های مختلف استفاده می‌کند، که در شکل ۲ نشان داده‌شده است. بنابراین، نتیجه دقت نهایی آزمایش‌های ما با استفاده از شبکه محاسبه می‌شود. هر اجرا دارای یک ترتیب ارائه تصادفی از نمونه‌های آموزشی است. وزن بهترین اجرا در مجموعه اعتبار سنجی بعداً برای ارزیابی مجموعه تست و محاسبه دقت نهایی استفاده می‌شود.

برای تأیید اینکه رویکرد پیشنهادی به‌خوبی سایر پایگاه‌های داده را مدیریت می‌کند، آزمایش‌هایی با پایگاه داده BU-3DFE، پایگاه داده JAFFE انجام شد. آزمایش‌ها با پایگاه‌های اطلاعاتی دیگر از همان رویکردی که پایگاه داده CKþ بود، پیروی می‌کنند. ارزیابی دقت با استفاده از پایگاه داده BU-3DFE و در پایگاه داده JAFFE محاسبه شد. در آزمایش‌های بین پایگاه داده، هیچ تصویری از پایگاه داده BU-3DFE یا پایگاه داده JAFFE در طول آموزش شبکه استفاده نشد.

جدول ۲ :تأثیر ترتیب در دقت

Presentation order	Accuracy (%)
1	88.18
2	86.36
3	86.36
4	86.36
5	88.18
6	84.55
7	85.45
8	84.55
9	87.27
10	88.18

جدول ۳ بهترین نتیجه به‌دست‌آمده (با استفاده از نرمال‌سازی‌ها و نمونه‌های مصنوعی) برای C6class و Cbin را نشان می‌دهد. همان‌طور که مشاهده می‌شود، رویکرد طبقه‌بندی کننده باینری دقت را افزایش می‌دهد. زیرا در این رویکرد، به‌جای استفاده از یک طبقه‌بندی، که در آن هر نمونه فقط یک شانس برای طبقه‌بندی مناسب دارد، می‌توان شش بار (یک بار برای هر طبقه‌بندی) به دست آورد. ما فکر می‌کنیم که طبقه‌بندی کننده شش کلاس (C6class) روش ارزیابی عادلانه‌تری است. بااین‌حال، رویکرد طبقه‌بندی Cbin زمانی که ما فقط به یک حالت علاقه‌مند هستیم، روشی مفید است. انحراف استاندارد گزارش‌شده در جدول بر اساس اجراهای همه آزمایش‌ها است.

جدول ۳ :دقت برای هر دو طبقه‌بندی کننده در پایگاه داده CKþ

Classifier	Angry (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)
C-6classE	93.33	100.00	96.00	98.55	84.52	99.20
C-binE	98.27	99.37	99.24	99.68	98.17	98.81
Average of C-6class: 96.76%±70.07						
Average of Cbin: 98.92%±70.02						

مقادیر پارامترهای آموزشی که به نتایج نشان داده‌شده در جدول ۳ می‌رسند در جدول ۴ نشان داده‌شده‌اند. از همان مقادیر پارامتر در آزمایش‌ها روی پایگاه‌های داده دیگر استفاده می‌شود.

جدول ۴ :پارامترهای آموزشی

Parameter	Value
Momentum	0.95
Learning rate	0.01
Epochs	10.00
Loss function	SoftmaxWithLoss
Gaussian standard deviation	3
Synthetic samples amount	70

با استفاده از نتیجه نشان داده‌شده در جدول ۳، ماتریس سردرگمی نشان داده‌شده در جدول ۵ برای طبقه‌بندی کننده کلاس شش ایجاد شد.

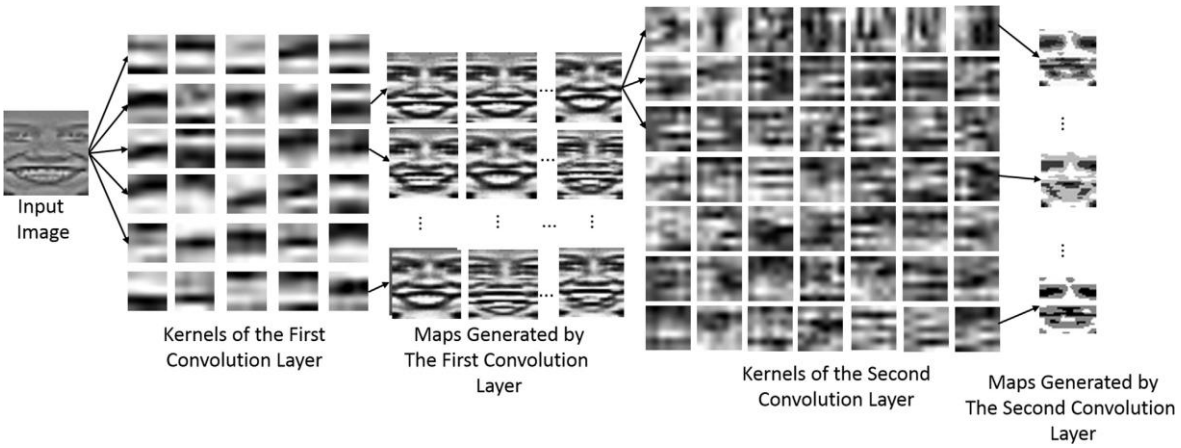
جدول ۵ :ماتریس درهم‌ریختگی با استفاده از نرمال‌سازی نمونه‌های مصنوعی در پایگاه داده CKþ

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	126	6	2	0	1	0
Disgust	0	177	0	0	0	0
Fear	0	0	72	0	3	0
Happy	3	0	0	204	0	0
Sad	3	0	1	0	71	9
Surprise	1	0	1	0	0	247

بر اساس نتایج طبقه‌بندی شش کلاسی، می‌توان به این اشاره کرد که حالات انزجار، خوشحالی و تعجب به میزان دقت بالاتر از ۹۸ درصد می‌رسند. درحالی‌که ابراز خشم و ترس به ترتیب حدود ۹۳ درصد و ۹۶ درصد بود. حالت غمگین با تنها ۸۴/۵۲ درصد کمترین میزان تشخیص را به دست می‌آورد. با نگاهی به ماتریس سردرگمی، حالت غمگین در اکثر مواقع با بیان تعجب اشتباه گرفته می‌شد. این نشان می‌دهد که ویژگی‌های این دو عبارت در فضای پیکسل به‌خوبی از هم جدا نشده‌اند، یعنی در برخی موارد بسیار شبیه به یکدیگر هستند. شکل ۱۱ نمونه‌هایی از طبقه‌بندی اشتباه را نشان می‌دهد.

شکل ۱۱: تشابه میان حالت غمگین و تعجب و اشتباه طبقه‌بندی کننده

شکل ۱۲ تصویری از هسته‌های آموخته‌شده و نقشه‌های تولیدشده برای هر لایه کانولوشن را نشان می‌دهد. در اولین لایه کانولوشن، تصویر ورودی توسط ۳۲ هسته پردازش می‌شود و ۳۲ نقشه خروجی تولید می‌کند. در لایه کانولوشن دوم، از ۶۴ هسته آموخته‌شده برای ایجاد نقشه‌های جدید برای هر یک از ۳۲ نقشه لایه قبلی استفاده می‌شود. هسته‌های نشان داده‌شده در شکل ۱۲ در آموزش با استفاده از پایگاه داده CKþ برای شش حالت اصلی آموزش داده شدند. همان‌طور که در شکل ۱۲ مشاهده می‌شود، پس از لایه کانولوشن دوم، نقشه‌های ایجادشده بر روی نواحی نزدیک چشم، دهان و بینی متمرکزشده‌اند. این مناطق برای تجزیه‌وتحلیل حالات چهره مهم‌تر هستند [۷۱].



شکل ۱۲: تصویری از هسته‌های آموخته‌شده و نقشه‌های تولیدشده برای هر لایه کانولوشن

۵– نتیجه‌گیری

در این مقاله، یک سیستم تشخیص حالت چهره را پیشنهاد می‌کنیم که از ترکیبی از روش‌های استاندارد، مانند شبکه عصبی کانولوشن و مراحل پیش‌پردازش تصویر استفاده می‌کند. آزمایش‌ها نشان داد که ترکیبی از روش‌های نرمال‌سازی به‌طور قابل‌توجهی دقت روش را بهبود می‌بخشد. همان‌طور که در نتایج نشان داده‌شده است، در مقایسه با روش‌های اخیر، که از پایگاه داده بیان چهره و روش تجربی استفاده می‌کنند، روش ما به نتایج بهتری دست می‌یابد و راه‌حل ساده‌تری ارائه می‌دهد. علاوه بر این آموزش زمان کمتری می‌برد و تشخیص آن در زمان واقعی انجام می‌شود. درنهایت، آزمایش‌های متقابل پایگاه داده نشان می‌دهد که رویکرد پیشنهادی در محیط‌های ناشناخته نیز کار می‌کند.

همان‌طور که در بخش ۱ توضیح داده شد، استفاده از شبکه‌های عصبی کانولوشنال باهدف کاهش نیاز به ویژگی‌های رمزگذاری شده دستی است. به‌جای مجموعه‌ای از ویژگی‌های انتخاب‌شده، ورودی آن می‌تواند تصاویر خام باشد. زیرا این مدل شبکه عصبی قادر به یادگیری مجموعه‌ای از ویژگی‌هایی است که به بهترین نحو طبقه‌بندی موردنظر را مدل می‌کند. برای انجام چنین یادگیری، شبکه‌های عصبی کانولوشن به حجم زیادی از داده‌ها نیاز دارند که ما آن را نداریم. این یک محدودیت از معماری‌های عمیق است که به مقدار زیادی پارامتر در طول آموزش نیاز دارند. برای رفع این مشکل (داده‌های محدود ما)، عملیات پیش‌پردازش روی تصاویر اعمال شد تا تغییرات بین تصاویر کاهش یابد و زیرمجموعه‌ای از ویژگی‌های یادگیری انتخاب شود، که نیاز به مقدار زیادی داده را کاهش می‌دهد. اگر مجموعه‌ای از تصاویر بهتر، با تنوع بیشتر و نمونه‌های بیشتر داشتیم، این عملیات پردازش اولیه نمی‌توانست برای دستیابی به‌دقت گزارش‌شده ضروری باشد و حتی اعتبارسنجی بین پایگاه‌های داده را می‌توانست بهبود دهد.

آزمایش‌های اولیه با معماری‌های عمیق‌تر، با حجم زیادی از داده‌ها انجام شد. در این آزمایش‌ها، یک شبکه عصبی کانولوشنال عمیق که توسط ۳۸ لایه تشکیل‌شده و با حدود ۹۸۲۸۰۰ تصویر از ۲۶۶۲ موضوع آموزش داده‌شده است، به‌طور خلاصه موردمطالعه قرار گرفت پیشنهادشده (توسط پرخی و همکاران [۸۳]). مدل از قبل آموزش‌دیده به‌عنوان یک استخراج‌کننده ویژگی استفاده شد که به‌عنوان ورودی یک شبکه عصبی دولایه ساده با پایگاه داده CKþ آموزش داده شد. در این آزمایش هیچ عملیات پیش‌پردازشی اعمال نشد. باوجود سادگی آزمایش، دقت به‌دست‌آمده در آزمایش‌های بین پایگاه‌های داده را افزایش دادند. این نتایج نشان می‌دهد که یک رویکرد یادگیری عمیق از این نوع می‌تواند راه بهتری برای تولید یک مدل متمایزکننده برای تشخیص حالات چهره باشد و به آن اجازه می‌دهد در پیش‌بینی‌های کنترل‌شده کار کند، که یکی از چالش‌های فعلی در این زمینه است.

به‌عنوان کار آینده، کاربرد روش استخراج ویژگی در مسائل دیگر بررسی خواهد شد. علاوه بر این، می‌خواهیم روش‌های یادگیری دیگری را برای افزایش استحکام روش در محیط‌های ناشناخته (مثلاً با شرایط مختلف نور و موارد دیگر) بررسی کنیم. همچنین، آزمایش‌های بیشتری با استفاده از توصیفگر صورت و استفاده از تکنیک‌های تنظیم دقیق، که هدف آن تنظیم یک شبکه عصبی عمیق آموزش‌دیده است، انجام می‌شود.

فهرست مراجع

[1] Y. Wu, H. Liu, H. Zha, Modeling facial expression space for recognition, in: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005 (IROS 2005), 2005, pp. 1968–1973.

[2] C. Darwin, The Expression of the Emotions in Man and Animals, CreateSpace Independent Publishing Platform, 2012.

[3] S.Z. Li, A.K. Jain, Handbook of Face Recognition, Springer Science & Business Media, Secaucus, NJ, USA, 2011.

[4] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn–Kanade dataset (CKþ): a complete dataset for action unit and emotion- specified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 94–101.

[5] M. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362.

[6] L. Yin, X. Wei, Y. Sun, J. Wang, M. Rosato, A 3d facial expression database for facial behavior research, in: 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Institute of Electrical & Electronics En- gineers (IEEE), Southampton, UK, 2006.

[7] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: 2014 IEEE Conference on Computer Vision and Pat- tern Recognition (CVPR), 2014, pp. 1805–1812.

[8] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.

[9] W. Liu, C. Song, Y. Wang, Facial expression recognition based on discriminative dictionary learning, in: 2012 21st International Conference on Pattern Re- cognition (ICPR), 2012, pp. 1839–1842.

[10] I. Song, H.-J. Kim, P.B. Jeon, Deep learning for real-time robust facial expression recognition on a smartphone, in: International Conference on Consumer Electronics (ICCE), Institute of Electrical & Electronics Engineers (IEEE), Las Vegas, NV, USA, 2014.

[11] P. Burkert, F. Trier, M.Z. Afzal, A. Dengel, M. Liwicki, Dexpression: Deep Con- volutional Neural Network for Expression Recognition, CoRR abs/1509.05371 (URL ⟨http://arxiv.org/abs/1509.05371⟩).

[12] M. Liu, S. Li, S. Shan, X. Chen, Au-inspired deep networks for facial expression feature learning, Neurocomputing 159 (2015) 126–136, http://dx.doi.org/ 10.1016/j.neucom.2015.02.011.

[13] G. Ali, M.A. Iqbal, T.-S. Choi, Boosted NNE collections for multicultural facial expression recognition, Pattern Recognit. 55 (2016) 14–27, http://dx.doi.org/ 10.1016/j.patcog.2016.01.032.

[14] Y.-H. Byeon, K.-C. Kwak, Facial expression recognition using 3d convolutional neural network. International Journal of Advanced Computer Science and Applications(IJACSA), 5 (2014).

[15] J.-J.J. Lien, T. Kanade, J. Cohn, C. Li, Detection, tracking, and classification of action units in facial expression, J. Robot. Auton. Syst. 31(3), 2000, 131-146.

[16] X. Fan, T. Tjahjadi, A spatial–temporal framework based on histogram of gra- dients and optical flow for facial expression recognition in video sequences, Pattern Recognit. 48 (11) (2015) 3407–3416.

[17] W. Zhang, Y. Zhang, L. Ma, J. Guan, S. Gong, Multimodal learning for facial expression recognition, Pattern Recognit. 48 (10) (2015) 3191–3202.

[18] C.-R. Chen, W.-S. Wong, C.-T. Chiu, A 0.64 mm real-time cascade face detection design based on reduced two-field extraction, IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 19 (11) (2011) 1937–1948.

[19] C. Garcia, M. Delakis, Convolutional face finder: a neural architecture for fast and robust face detection, IEEE Trans. Pattern Anal. Mach. Intell. 26 (11) (2004) 1408–1423, http://dx.doi.org/10.1109/TPAMI.2004.97.

[20] Z. Zhang, D. Yi, Z. Lei, S. Li, Regularized transfer boosting for face detection across spectrum, IEEE Signal Process. Lett. 19 (3) (2012) 131–134.

[21] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Re- cognizing facial expression: machine learning and application to spontaneous behavior, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005), vol. 2, 2005, pp. 568–573.

[22] P. Liu, M. Reale, L. Yin, 3d head pose estimation based on scene flow and generic head model, in: 2012 IEEE International Conference on Multimedia and Expo (ICME), 2012, pp. 794–799.

[23] W.W. Kim, S. Park, J. Hwang, S. Lee, Automatic head pose estimation from a single camera using projective geometry, in: 2011 8th International Conference on In- formation, Communications and Signal Processing (ICICS), 2011, pp. 1–5.

[24] M. Demirkus, D. Precup, J. Clark, T. Arbel, Multi-layer temporal graphical model for head pose estimation in real-world videos, in: 2014 IEEE Interna- tional Conference on Image Processing (ICIP), 2014, pp. 3392–3396.

[25] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry- based and gabor-wavelets-based facial expression recognition using multi- layer perceptron, in: 1998 Proceedings of the Third IEEE International Con- ference on Automatic Face and Gesture Recognition, 1998, pp. 454–459.

[26] P. Yang, Q. Liu, D. Metaxas, Boosting coded dynamic features for facial action units and facial expression recognition, in: IEEE Conference on Computer Vi- sion and Pattern Recognition, 2007 (CVPR'07), 2007, pp. 1–6.

[27] S. Jain, C. Hu, J. Aggarwal, Facial expression recognition with temporal mod- eling of shapes, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, pp. 1642–1649, http://dx.doi.org/10.1109/ ICCVW.2011.6130446.

[28] Y. Lin, M. Song, D.T.P. Quynh, Y. He, C. Chen, Sparse coding for flexible, robust 3d facial-expression synthesis, IEEE Comput. Graph. Appl. 32 (2) (2012) 76–88.

[29] S. Rifai, Y. Bengio, A. Courville, P. Vincent, M. Mirza, Disentangling factors of variation for facial expression recognition, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), Computer Vision – ECCV 2012, Lecture Notes in Computer Science, vol. 7577, Springer, Berlin Heidelberg, 2012, pp. 808–822.

[30] B. Fasel, Robust face analysis using convolutional neural networks, in: Pro- ceedings of the 16th International Conference on Pattern Recognition, 2002, vol. 2, 2002, pp. 40–43.

[31] F. Beat, Head-pose invariant facial expression recognition using convolutional neural networks, in: Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces, 2002, 2002, pp. 529–534.

[32] M. Matsugu, K. Mori, Y. Mitari, Y. Kaneda, Subject independent facial expres- sion recognition with robust face detection using a convolutional neural network, Neural Netw.: Off. J. Int. Neural Netw. Soc. 16 (5) (2003) 555–559.

[33] Y. Bengio, Y. LeCun, Scaling learning algorithms towards AI, in: L. Bottou, O. Chapelle, D. DeCoste, J. Weston (Eds.), Large-Scale Kernel Machines, MIT Press, Cambridge, Massachusetts, USA, 2007 (URL ⟨http://yann.lecun.com/exdb/pub lis/pdf/bengio-lecun-07.pdf⟩).

[34] P.E. Utgoff, D.J. Stracuzzi, Many-layered learning, Neural Comput. 14 (10) (2002) 2497–2529.

[35] Y. Bengio, I.J. Goodfellow, A. Courville, Deep Learning, MIT Press, Cambridge, Massachusetts, USA, 2015.

[36] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based Learn. Appl. Doc. Recognit. 86 (11) (1998) 2278–2324, http://dx.doi.org/10.1109/5.726791.

[37] D.C. Cirean, U. Meier, J. Masci, L.M. Gambardella, J. Schmidhuber, Flexible, high performance convolutional neural networks for image classification, in: Pro- ceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11), vol. 2, AAAI Press, Barcelona, Catalonia, Spain, 2011, pp. 1237–1242.

[38] P. Zhao-yi, W. Zhi-qiang, Z. Yu, Application of mean shift algorithm in real- time facial expression recognition, in: International Symposium on Computer Network and Multimedia Technology, 2009 (CNMT 2009), 2009, pp. 1–4.

[39] H.Y. Patil, A.G. Kothari, K.M. Bhurchandi, Expression invariant face recognition using local binary patterns and contourlet transform, Opt.-Int. J. Light Electron Opt. 127 (5) (2016) 2670–2678, http://dx.doi.org/10.1016/j.ijleo.2015.11.187.

[40] J.Y.R. Cornejo, H. Pedrini, F. Florez-Revuelta, Facial expression recognition with occlusions based on geometric representation, in: Progress in Pattern Recogni- tion, Image Analysis, Computer Vision, and Applications: Proceedings of the 20th Iberoamerican Congress (CIARP 2015), Montevideo, Uruguay, November 9–12, 2015, Springer International Publishing, Cham, 2015, pp. 263–270.

[41] Z. Wang, Q. Ruan, G. An, Facial expression recognition using sparse local fisher discriminant analysis, Neurocomputing 174 (Part B) (2016) 756–766, http://dx. doi.org/10.1016/j.neucom.2015.09.083.

[42] S. Arivazhagan, R.A. Priyadharshini, S. Sowmiya, Facial expression recognition based on local directional number pattern and anfis classifier, in: 2014 Inter- national Conference on Communication and Network Technologies (ICCNT), 2014, pp. 62–67 (http://dx.doi.org/10.1109/CNT.2014.7062726).

[43] A.R. Rivera, J.R. Castillo, O.O. Chae, Local directional number pattern for face analysis: face and expression recognition, IEEE Trans. Image Process. 22 (5) (2013) 1740–1752.

[44] T.H. Zavaschi, A.S. Britto, L.E. Oliveira, A.L. Koerich, Fusion of feature sets and classifiers for facial expression recognition, Expert Syst. Appl. 40 (2) (2013) 646–655, http://dx.doi.org/10.1016/j.eswa.2012.07.074.

[45] A.T. Lopes, E. de Aguiar, T.O. Santos, A facial expression recognition system using convolutional networks, in: 2015 28th SIBGRAPI Conference on Gra- phics, Patterns and Images, Institute of Electrical & Electronics Engineers (IEEE), Salvador, Bahia, Brasil, 2015.

[46] S. Demyanov, J. Bailey, R. Kotagiri, C. Leckie, Invariant Backpropagation: How To Train a Transformation-Invariant Neural Network (arXiv:1502.04434[cs, stat]).

[47] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding (ar Xiv:1408.5093).

[48] C.-D. Caleanu, Face expression recognition: a brief overview of the last decade, in: 2013 IEEE 8th International Symposium on Applied Computational In- telligence and Informatics (SACI), 2013, pp. 157–161.

[49] M.K.A.E. Meguid, M.D. Levine, Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers, IEEE Trans. Affect. Comput. 5 (2) (2014) 141–154, http://dx.doi.org/10.1109/TAFFC.2014.2317711.

[50] C. Turan, K. M. Lam, Region-based feature fusion for facial-expression re- cognition, in: 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 5966–5970 (http://dx.doi.org/10.1109/ICIP.2014.7026204).

[51] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, 1998, pp. 200–205.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Drop- out: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[53] J.M. Girard, J.F. Cohn, L.A. Jeni, S. Lucey, F.D. la Torre, How much training data for facial action unit detection?, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, 2015, pp. 1–8 (http://dx.doi.org/10.1109/FG.2015.7163106).

[54] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proceedings of the 3rd International Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, 2010, p. 65.

[55] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Work- shops), IEEE, Barcelona, Catalonia, Spain, 2011, pp. 2106–2112.

[56] P. Simard, D. Steinkraus, J.C. Platt, Best practices for convolutional neural networks applied to visual document analysis, in: 2003 Proceedings of the Seventh International Conference on Document Analysis and Recogni- tion, 2003, pp. 958–963.

[57] J.-I. Choi, C.-W. La, P.-K. Rhee, Y.-L. Bae, Face and eye location algorithms for visual user interface, in: Proceedings of First Signal Processing Society Workshop on Multimedia Signal Processing, Institute of Electrical & Electro- nics Engineers (IEEE), Princeton, NJ, USA, 1997.

[58] G. Li, X. Cai, X. Li, Y. Liu, An efficient face normalization algorithm based on eyes detection, in: 2006 IEEE/RSJ International Conference on Intelligent Ro- bots and Systems, Institute of Electrical & Electronics Engineers (IEEE), Beijing, China, 2006.

[59] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vision. 91 (2) (2010) 200–215.

[60] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE Con- ference on Computer Vision and Pattern Recognition, 2013, pp. 3444–3451.

[61] S. Cheng, A. Asthana, S. Zafeiriou, J. Shen, M. Pantic, Real-time generic face tracking in the wild with cuda, in: Proceedings of the 5th ACM Multimedia Systems Conference, ACM, Singapore, Singapore 2014, pp. 148–151.

[62] B.A. Wandell, Foundations of Vision, 1st ed., Sinauer Associates Inc, Sunder- land, Mass, 1995.

[63] L. Bottou, Stochastic Gradient Descent Tricks, Springer, New York, NY, USA. 2012.

[64] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10), Society for Artificial Intelligence and Statistics, Sardinia, Italy, 2010.

[65] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: G.J. Gordon, D.B. Dunson (Eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11), vol. 15, 2011, pp. 315–323.

[66] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, 1978.

[67] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, H. Lin, Facial expression recognition using radial encoding of local gabor features and classifier synthesis, Pattern Recognit. 45 (1) (2012) 80–91, http://dx.doi.org/10.1016/j.patcog.2011.05.006.

[68] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D. Metaxas, Learning active facial patches for expression analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2562–2569.

[69] S.H. Lee, W.J. Baddar, Y.M. Ro, Collaborative expression representation using peak expression and intra class variation face images for practical subject- independent emotion recognition in videos, Pattern Recognit. 54 (2016) 52–67, http://dx.doi.org/10.1016/j.patcog.2015.12.016.

[70] F.D. la Torre, W.S. Chu, X. Xiong, F. Vicente, X. Ding, J. Cohn, Intraface, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, 2015, pp. 1–8 (http://dx.doi.org/10.1109/FG. 2015.7163082).

[71] J. Cohn A. Zlochower, A Computerized Analysis of Facial Expression: Feasibility of Automated Discrimination, vol. 2. American Psychological Society, 1995, p. 6.

[72] M. Xue, A. Mian, W. Liu, L. Li, Fully automatic 3d facial expression recognition using local depth features, in: IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1096–1103 (http://dx.doi.org/10.1109/WACV.2014. 6835736).

[73] T. Sha, M. Song, J. Bu, C. Chen, D. Tao, Feature level analysis for 3d facial ex- pression recognition, Neurocomputing 74 (12–13) (2011) 2135–2141, http://dx. doi.org/10.1016/j.neucom.2011.01.008.

[74] A. Maalej, B.B. Amor, M. Daoudi, A. Srivastava, S. Berretti, Shape analysis of local facial patches for 3d facial expression recognition, Pattern Recognit. 44 (8) (2011) 1581–1589, http://dx.doi.org/10.1016/j.patcog.2011.02.012.

[76] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1749– 1756.

[77] D. Mery, K. Bowyer, Automatic facial attribute analysis via adaptive sparse representation of random patches, Pattern Recognit. Lett. 68 (Part 2) (2015) 260–269 (Special Issue on “Soft Biometrics”,.

[78] M.H. Siddiqi, R. Ali, A.M. Khan, Y.T. Park, S. Lee, Human facial expression re- cognition using stepwise linear discriminant analysis and hidden conditional random fields, IEEE Trans. Image Process. 24 (4) (2015) 1386–1398, http://dx. doi.org/10.1109/TIP.2015.2405346.

[79] A. Zafer, R. Nawaz, J. Iqbal, Face recognition with expression variation via ro- bust ncc, in: 2013 IEEE 9th International Conference on Emerging Technolo- gies (ICET), 2013, pp. 1–5 (http://dx.doi.org/10.1109/ICET.2013.6743520).

[80] M.H. Siddiqi, R. Ali, A.M. Khan, E.S. Kim, G.J. Kim, S. Lee, Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection, Multimed. Syst. 21 (6) (2014) 541–555, http://dx.doi.org/10.1007/s00530-014-0400-2.

[81] M.H. Siddiqi, R. Ali, M. Idris, A.M. Khan, E.S. Kim, M.C. Whang, S. Lee, Human facial expression recognition using curvelet feature extraction and normalized mutual information feature selection, Multimed. Tools Appl. 75 (2) (2014) 935–959, http://dx.doi.org/10.1007/s11042-014-2333-3.

[82] T. Kanade, Y. Tian, J.F. Cohn, Comprehensive database for facial expression analysis, in: Proceedings of the Fourth IEEE International Conference on Au- tomatic Face and Gesture Recognition 2000 (FG'00), IEEE Computer Society, Washington, DC, USA, 2000, p. 46.

[83] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Ma- chine Vision Conference, 2015, 46-53.